CODEN: JASMAN

The Journal of the Acoustical Society of America

ISSN: 0001-4966

Vol. 119, No. 3 March 2006

(Continued)

CONTENTS—*Continued from preceding page*

(Continued)

CONTENTS—*Continued from preceding page*

(Continued)

CONTENTS—*Continued from preceding page*

SPEECH PROCESSING AND COMMUNICATION SYSTEMS [72]

MUSIC AND MUSICAL INSTRUMENTS [75]

BIOACOUSTICS [80]

ERRATA

JASA EXPRESS LETTERS

CUMULATIVE AUTHOR INDEX

The effect of source amplitude and phase in matched field source localization

Zoi-Heleni Michalopoulou
Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, New Jersey 07102
michalop@njit.edu

Abstract: Traditional matched field processing is typically employed for localization of a source with unknown amplitude and phase at transmission. The spectral information of the source is estimated via maximum likelihood and is used for localization with the Bartlett processor. Maximum *a posteriori* estimation of the spectrum of the source is here proposed along with the estimation of source location and noise variance, where the joint posterior probability distribution of all parameters is computed with Gibbs sampling. It is shown that the proposed method has a higher probability of correct localization than the traditional approach.

PACS numbers: 43.60.Jn, 43.60.Kx, 43.30.Wi [JVC]
Date Received: November 13, 2005 **Date Accepted:** December 8, 2005

1. Introduction

Matched field processing (MFP) is often employed for source localization in underwater environments.[1,2] Matching the true acoustic field to field replicas generated with a sound propagation model, MFP produces estimates that generate the best match between true and replica fields according to some specified matching criterion. The most widely used MFP technique is the Bartlett or linear processor, which calculates inner products between the received data and normalized replica fields.

In this work the derivation of the Bartlett processor using a simple, spatially white, Gaussian noise environment is reviewed, with emphasis on the role of the source spectrum (and uncertainty characterizing source spectrum information) on source localization. In contrast to Ref. 3, where the source spectrum was deemed useful only when the exact times at which measurements are made are available, we show that accounting for uncertainty in the source spectrum is useful under more general circumstances.

The structure of the paper is as follows: Sec. II discusses the derivation of the Bartlett estimator. Section III presents a maximum a posteriori (MAP) estimator for source localization including source amplitude and phase and noise variance as unknowns; the posterior probability distribution of all unknowns is estimated using Gibbs sampling. Section IV demonstrates a comparison between the proposed MAP processor and a simple Bartlett processor for synthetic data. Conclusions are presented in Sec. V.

2. Maximum likelihood source localization

In a typical underwater source localization problem, measurements of the acoustic field at frequency f are acquired at a set of L spatially (here, vertically) separated hydrophones. In the presence of spatially white, Gaussian noise, measurements $\mathbf{X}$ can be written as

$$\mathbf{X} = \alpha \mathbf{G}(r, z_s) + \mathbf{W}, \tag{1}$$

where $\mathbf{G}(r, z_s)$ is the Green's function obtained by solving the Helmholtz equation,[4] or the theoretically expected field for a given source location and underwater environment for a source at frequency f with amplitude 1 and phase 0; it is assumed that the propagation environment is exactly known with the only unknowns necessary to determine $\mathbf{G}$ being source range r and depth z_s. Factor α is a complex number containing the source amplitude and phase at trans-

mission, and **W** is additive noise drawn from a complex Gaussian distribution with 0 mean and diagonal spatial covariance matrix $\Sigma=2\sigma^2 I_{L\times L}$.

For the spatially white Gaussian noise environment selected here, the received field **X** can be modeled with a Gaussian distribution,

$$p(\mathbf{X}|r,z_s,\alpha)=\frac{1}{\sigma^{2L}(2\pi)^L}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{X}-\alpha\mathbf{G}(r,z_s))^*(\mathbf{X}-\alpha\mathbf{G}(r,z_s))\right). \quad (2)$$

Once data **X** are observed, Eq. (2) provides the likelihood function for α, r, and z_s,

$$l(r,z_s,\alpha,\sigma^2|\mathbf{X})\propto\exp\left(-\frac{1}{2\sigma^2}(\mathbf{X}-\alpha\mathbf{G}(r,z_s))^*(\mathbf{X}-\alpha\mathbf{G}(r,z_s))\right). \quad (3)$$

We are interested in obtaining the maximum likelihood estimates of r and z_s; factors α and σ^2 are nuisance parameters. Typically in MFP, the maximum likelihood estimate $\hat{\alpha}$ of α is found, and it is then placed instead of α in Eq. (3).[5] The maximum likelihood estimate $\hat{\alpha}$ of α is

$$\hat{\alpha}=\frac{\mathbf{G}^*(r,z_s)\mathbf{X}}{\|\mathbf{G}(r,z_s)\|^2}. \quad (4)$$

Once $\hat{\alpha}$ is placed in Eq. (3), the likelihood can be further maximized over σ^2.[5] The likelihood for range and source depth then becomes

$$l_{r,z_s}(r,z_s|\mathbf{X})\propto\left(\|\mathbf{X}\|^2-\frac{\mathbf{G}^*\mathbf{X}\mathbf{X}^*\mathbf{G}}{\|\mathbf{G}\|^2}\right)^{-L}. \quad (5)$$

Maximization of l_{r,z_s} over r and z_s is equivalent to maximization of P over r and z_s, where

$$P(r,z_s)=\frac{\mathbf{G}^*\mathbf{X}\mathbf{X}^*\mathbf{G}}{\|\mathbf{G}\|^2}. \quad (6)$$

Maximizing quantity P (typically called an ambiguity surface) is referred to as the Bartlett or linear matched-field processor.

3. Gibbs sampling for matched field processing

An alternative way of looking at the same estimation problem is through maximization of posterior probability distributions. Estimation of unknown parameters can be obtained through the maximization of the posterior probability distribution of these parameters given the observed data and prior knowledge. Assuming that **q** is a vector containing all unknown parameters, this posterior probability distribution is given by

$$p(\mathbf{q}|\mathbf{X})=\frac{l(\mathbf{q}|\mathbf{X})p(\mathbf{q})}{p(\mathbf{X})}. \quad (7)$$

Distribution $p(\mathbf{q})$ contains all prior knowledge on **q** and $p(\mathbf{X})$ is a constant with respect to **q**. Quantity $l(\mathbf{q}|\mathbf{X})$ is the likelihood for parameters contained in vector **q** after observing data **X**.

Here, $\mathbf{q}=[r\ z_s\ \alpha\ \sigma^2]^T$. Assuming no specific information on the source spectrum, the prior distribution on α reflects that its real and imaginary parts can vary between $-\infty$ and ∞; specifically we choose the prior distribution of α,

$$p(\alpha)=C,\quad -\infty<\mathrm{Re}(\alpha),\ \mathrm{Im}(\alpha)<\infty, \quad (8)$$

where C is a constant. This modeling of prior information is more general than the Rayleigh and uniform modeling of prior information on amplitude and phase respectively[6] (it does not involve knowledge or assumptions on the Rayleigh parameter).

For the unknown source range and depth, we select a uniform prior in intervals $[r_1, r_2]$ for range r and $[z_{s1}, z_{s2}]$ for depth z_s,

$$p(r) = \frac{1}{r_2 - r_1}, \tag{9}$$

and

$$p(z_s) = \frac{1}{z_{s2} - z_{s1}}. \tag{10}$$

Another unknown quantity in Eq. (3) is variance σ^2. As in Ref. 7, we select a noninformative prior distribution for variance,

$$p(\sigma^2) = \frac{1}{\sigma^2}, \quad 0 < \sigma < \infty. \tag{11}$$

Combining Eqs. (3), (8), (9), (10), and (11), Eq. (7) becomes

$$p(r, z_s, \alpha, \sigma^2 | \mathbf{X}) = M \frac{1}{\sigma^{2L+2}} \exp((-1/(2\sigma^2))(\mathbf{X} - \alpha\mathbf{G})^*(\mathbf{X} - \alpha\mathbf{G})), \tag{12}$$

where M is a constant

Following the Bayesian paradigm, the focus is on calculating the full joint posterior distribution of all unknown parameters; integrating then over α and σ^2, we find the marginal posterior distribution of r and z_s, the main mode of which will yield the source location estimates. Distribution, $p(r, z_s | \mathbf{X})$ is given by

$$p(r, z_s | \mathbf{X}) = \int_{\alpha} \int_{\sigma^2} p(r, z_s, \alpha, \sigma^2 | \mathbf{X}) d\sigma^2 d\alpha. \tag{13}$$

This integration over α and σ^2 is in contrast to the action taken in Sec. II, where point (maximum likelihood) estimates of α and σ^2 were introduced in the likelihood function, which was then maximized for r and z_s. In the place of the point estimates of α and σ^2, we now take into account the uncertainty on α and σ^2 as well.

4. Gibbs sampling for source localization and source spectrum estimation

Gibbs sampling[8–11] allows us to estimate $p(r, z_s | \mathbf{X})$ and its mode in an efficient way without exhaustively calculating $p(r, z_s, \alpha, \sigma^2 | \mathbf{X})$ and integrating over α and σ^2. For the implementation of the Gibbs sampler, we need to identify the conditional marginal posterior distributions for r, z_s, α, and σ^2.

For the marginal posterior distribution of α we need to fix r, z_s, and σ^2 in Eq. (12),

$$p(\alpha | r, z_s, \sigma^2, \mathbf{X}) = M_\alpha \exp((-\|\mathbf{G}\|^2/(2\sigma^2))(\alpha - \mathbf{G} * \mathbf{X}/\|\mathbf{G}\|^2)^*(\alpha - \mathbf{G} * \mathbf{X}/\|\mathbf{G}\|^2)), \tag{14}$$

where M_α is a constant with respect to α. Complex variable α is, thus, normally distributed with mean $\mathbf{G}^*\mathbf{X}/\|\mathbf{G}\|^2$ and variance $2\sigma^2/\|\mathbf{G}\|^2$.

The conditional posterior distribution for the variance is identified as an inverse χ^2 distribution,

$$p(\sigma^2 | \, r, z_s, \alpha, \mathbf{X})) = \frac{1}{\sigma^{2L+2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{X} - \alpha\mathbf{G}(r, z_s)\|^2\right). \tag{15}$$

Samples from such a distribution can be drawn readily.

The conditional distribution of r and z_s is as follows:

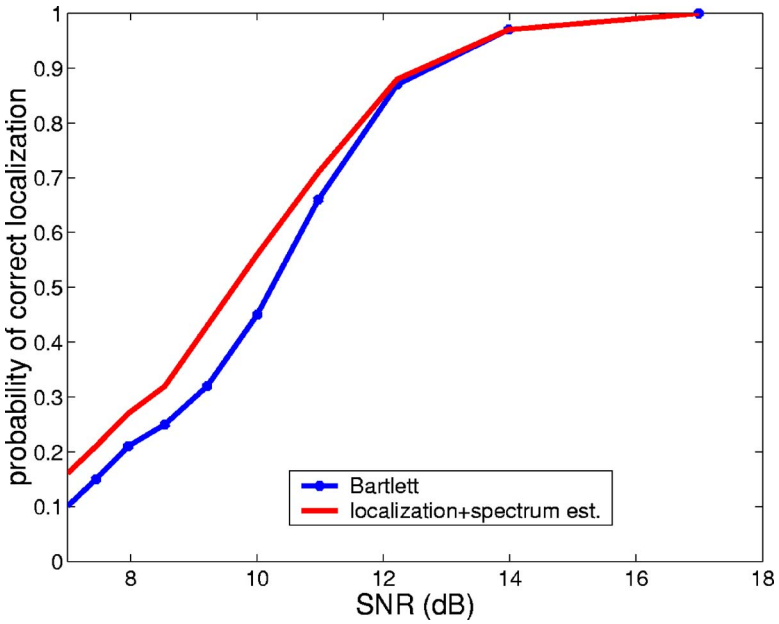


Fig. 1. (Color online) Probability of correct localization vs SNR at 600 Hz.

$$p(r,z_s|\ \alpha,\sigma^2,\mathbf{X})) = K\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{X}-\alpha\mathbf{G}(r,z_s)\|^2\right), \tag{16}$$

where K is a constant. The distribution is evaluated on a grid for r and z_s similarly to the implementation in Ref. 12 and following the griddy Gibbs modeling of Ref. 13, because it is not a closed from.

Gibbs sampling begins with a set of randomly chosen initial conditions for all unknown parameters (r, z_s, α, and σ^2). The process as implemented here first draws a sample from the conditional distribution of range r and source depth z_s; this two-dimensional sample contains the new, updated values of range and depth for the first iteration. Subsequently, a sample is drawn from the Gaussian marginal conditional posterior of α [Eq. (14)]. Lastly, a sample is drawn for the variance from the inverse χ^2 distribution of Eq. (15), completing the first iteration. For a large number of iterations, the obtained sample sequences eventually converge to the true joint posterior distribution of r, z_s, α, and σ^2.[10,8,9]

5. Performance evaluation in source localization

To evaluate the proposed method, synthetic data were generated simulating sound propagating in a shallow water environment. Field measurements at a vertical array of 24 phones were constructed for a frequency of 600 Hz. Signal to noise ratios (SNRs) between 7 and 14 dB were considered and one hundred noisy realizations were generated per SNR. Source localization was then performed with a simple Bartlett processor and with the proposed estimator; probabilities of correct source localization were estimated for the two processors. The true range and depth were 2 km and 34 m, respectively. Correct localization was defined as a result with a range estimate between 1.8 and 2.2 km and a source depth estimate between 28 and 40 m. The search interval for range was between 0 and 5 km with a 0.02 km spacing; the interval for depth was between 0 and 72 m with a 2 m spacing.

Figure 1 shows probability of correct localization vs SNR for the two processors. Consistently, the new processor has a higher probability of correct localization than the simple

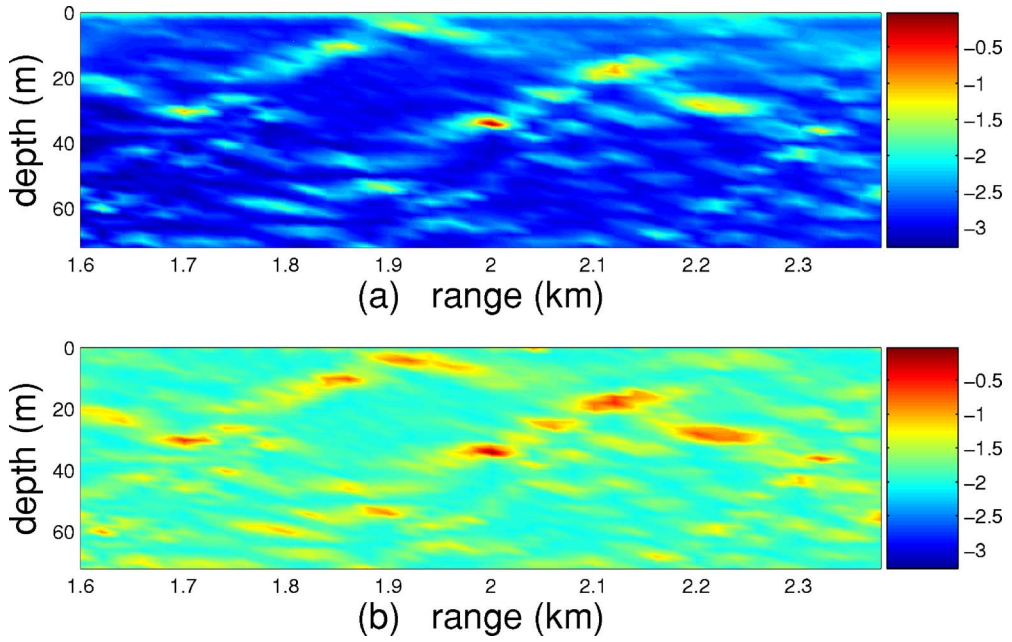


Fig. 2. (Color online) Posterior probability distribution (a) accounting for uncertainty in α and σ^2 and (b) employing the maximum likelihood estimates of α and σ^2.

Bartlett processor. The improvement in performance is negligible in high SNRs, where the maximum likelihood estimate of the source spectrum is accurate. The gap in performance increases as the SNR reaches lower values.

To highlight further the performance improvement with the new processor, we calculated the posterior distribution of range r and depth z_s using the proposed method (accounting for the uncertainty in α and σ^2) and using the conventional approach where the maximum likelihood estimates of α and σ^2 were used; the calculations were performed for an SNR of 8 dB and the computed distributions are shown in Fig. 2. The probability distribution calculated after integration over α and σ^2 has taken place has lower sidelobes than the distribution that uses point estimates, resulting in fewer mislocalizations, explaining the curves of Fig. 1.

6. Conclusions

A new matched field approach was presented, incorporating source spectrum as an unknown in the source localization process. The method provides improved localization performance as the noise in the environment increases degrading the point estimate for source spectrum typically employed in conventional matched-field processing. In real data applications, when modeling is approximate and the maximum likelihood estimate of the spectrum could be less reliable, it is anticipated that the new method will be of great value .

Acknowledgments

This work was supported by the Office of Naval Research through Grant Nos. N00014-00-1-0051 and N00014-97-1-0600. The author is grateful for computing support from the NSF MRI Program.

References and links

[1]A. Tolstoy, *Matched Field Processing for Underwater Acoustics* (World Scientific, Singapore, 1993).
[2]A. Baggeroer, W. Kuperman, and H. Schmidt, "Matched field processing: Source localization in correlated noise as an optimum parameters estimation problem," J. Acoust. Soc. Am. **83**, 571–587 (1988).

[3]Z.-H. Michalopoulou, "Gibbs sampling optimization in underwater sound problems," in *Proceedings of OCEANS 2001*, Hawaii, Vol. **2**, 782–785.
[4]F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, *Computational Ocean Acoustics*, (AIP Press, New York, 1994).
[5]P. Gerstoft and C. Mecklenbräuker, "Ocean acoustic inversion with estimation of a posteriori probability distributions," J. Acoust. Soc. Am. **104**(2), 808–819 (1998).
[6]A. Richardson and L. Nolte, "*A posteriori* probability source localization in an uncertain sound speed, deep ocean environment," J. Acoust. Soc. Am. **89**(5), 2280–2284 (1991).
[7]Z.-H. Michalopoulou and M. Picarelli, "Gibbs sampling for time-delay and amplitude estimation in underwater acoustics," J. Acoust. Soc. Am. **117**, 799–808 (2005).
[8]S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-6**, 721–741 (1984).
[9]A. E. Gelfand and A. F. Smith, "Sampling-based approaches to calculating marginal densities," J. Am. Stat. Assoc. **85**, 398–409 (1990).
[10]W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, 1st ed. (Chapman and Hall, New York, 1996).
[11]S. Dosso, "Quantifying uncertainty in matched field inversion. I A fast Gibbs sampler approach," J. Acoust. Soc. Am. **111**(1), 129–142 (2002).
[12]Z.-H. Michalopoulou and M. Picarelli, "Gibbs sampling for time delay and amplitude estimation in underwater acoustics," J. Acoust. Soc. Am. **117**, 799–808 (2005).
[13]M. A. Tanner, *Lecture Notes in Statistics: Tools for Statistical Inference (Observed Data and Data Augmentation Methods)*, 2nd ed. (Springer Verlag, Berlin, 1992), Vol. **67**.

Electromechanical emulation of active vibratory systems

Wenyuan Chen[a] **and Pierre E. Dupont**
Department of Aerospace and Mechanical Engineering, College of Engineering, Boston University, Boston, Massachusetts 02215
wychen@alum.bu.edu; pierre@bu.edu

Abstract: The design of a simple electromechanical system which is dynamically equivalent to an active vibratory system is studied in this paper. A two-step process is presented in which a passive vibratory system is first obtained, which is then modified in the second step to achieve active equivalence. Implementation of the active emulation step is achieved by closed-loop control of electromechanical shakers attached to the passive system and driven so as to generate the appropriate vibrations at the mounting locations of the active vibratory system. Experimental examples are used to demonstrate the effectiveness of this design process.

PACS numbers: 43.40.At, 43.40.Vn [JGM]
Date Received: July 21, 2005 **Date Accepted:** December 27, 2005

1. Introduction

The electromechanical emulation problem concerns the design and fabrication of dynamically equivalent models for complicated structures. This problem arises in the scaled vibration testing of novel superstructures or foundations for ships, submarines, and space structures. Equipment and machinery attached to the superstructure can both absorb and generate vibrational energy and so testing must also include its dynamics. In such testing, scaled models of the superstructure are often easy to construct. The scaled modeling of electronic equipment and power plant machinery is not, however, as straightforward. In these situations, an economical alternative to detailed scaling is the use of emulators—simple electromechanical systems which are dynamically equivalent only at the mounting locations where the equipment is attached to the superstructure.

The electromechanical filter design problem is similar except that, in this case, one starts with an actual machine instead of a desired frequency response function. Experiment or finite element simulation is used to extract the mounting location dynamic response, usually in the form of accelerance. Subsequently, mathematical models are fit to the data. These models are of a form that can be directly interpreted as an interconnection of easily realizable electromechanical elements.

A variety of techniques have been developed for passive emulation which specifically address the identification of, or conversion to, a model in realizable form. See, for example, Refs. 1–4.

Active emulation, which is the topic of this paper, involves reproducing the internally-generated vibrations which are transmitted to the superstructure. These vibrations are caused by such elements as unbalanced rotors and are mainly determined by the amplitudes and directions of the internal exciting forces, the transmission paths from those forces to the mounting locations, and the dynamic interaction between active machinery and its mounting structures.

---

[a]Presently at AC Technology Corporation, Member of the Lenze Group, 630 Douglas Street, Uxbridge, Massachusetts 01569.

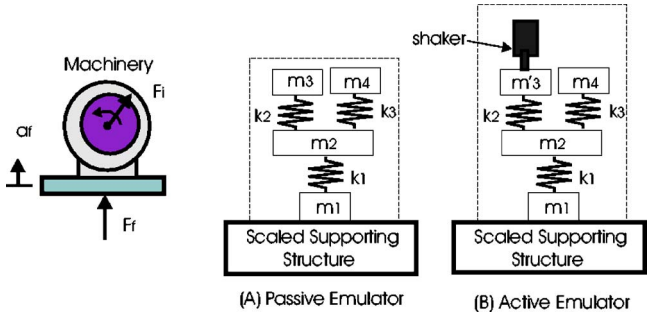


Fig. 1. (Color online) Two-step machinery emulator design procedure.

Active emulation has apparently not been addressed in the literature. The approach presented here is to modify a passive emulator to achieve active emulation. The method is described in the next section and is followed by an experimental validation.

2. Solution approach

The active emulation problem is solved by a two-step procedure as illustrated in Fig. 1. In the first step, a passive mechanical emulator is designed to match the passive dynamics of the machinery. The passive emulator must possess the rigid body modes and passive frequency responses of the original machinery within the frequency range of interest. It is assumed that this model is obtained by one of the existing methods in the literature. In the second step, this passive emulator is modified by adding active elements, such as shakers, to produce the active vibrations at the foundation mounting locations.

This approach assumes linearity of the models, which is typically valid for above-mount machinery models, i.e., models which are defined to exclude elastomeric mounts. Modeling of elastomeric mounts is treated elsewhere, e.g., Ref. 5. Using superposition, the total acceleration, a_f, at the mounting locations due to forces from both the superstructure and from internal moving components is given by

$$a_f = a_f^p + a_f^a = A_{ff} f_f + A_{fi} f_i. \tag{1}$$

The dimension of the vector a_f is the product of the number of mounting locations and the degrees of freedom per mounting location. The passive and active accelerations are denoted a_f^p and a_f^a, respectively. A_{ff} is the drive-point accelerance matrix of the passive (nonoperating) machinery and f_f is the vector of forces exerted by the superstructure on the passive machinery. Experimentally, A_{ff} can be estimated by suspending the machinery and driving the mounting locations with shakers through impedance heads. A_{fi} is the transfer accelerance matrix between the active internal forces f_i and the mounting locations. While f_i cannot usually be measured, a_f^a can be measured, for example, by suspending the operating machinery and measuring the accelerations of the freely-suspended mounting locations.

The passive emulator is designed to reproduce A_{ff} at its mounting locations. The active accelerations, a_f^a, are reproduced by attaching shakers to the passive emulator and driving them appropriately. Usually, at least one shaker should be used to reproduce motion in each coordinate direction. By using two emulating shakers, rotational accelerations may be generated.

In the approach adopted here, emulating shakers are commanded to produce the force corresponding to the mounting location acceleration occurring when there is no external loading. As given in Eq. (1), this technique provides for total mounting location acceleration to be a linear combination of the effects of internal and external forcing. It is assumed that external excitation does not affect the equivalent internal forcing of operating machinery.

This approach to active emulator design involves three aspects: selection of the emulating shakers, selection of their mounting locations in the passive emulator, and design of a controller for the shakers. These are discussed below.

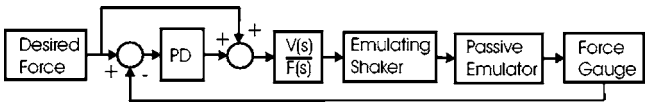


Fig. 2. Feedback control of the emulating shaker.

2.1 Selection of emulating shakers and their mounting locations

An emulating shaker must satisfy both passive and active matching requirements. Since the mass and rotational inertia of a shaker are appreciable, the mass element of the passive emulator to which the shaker is attached must be modified accordingly. For the single degree of freedom example in Fig. 1, the mass m_3' must satisfy

$$m_3' = m_3 - m_s > 0, \tag{2}$$

where m_s is the emulating shaker mass. In general, approximating the nondriven shaker as a rigid mass, the combined shaker and emulator mass element should meet all passive rigid body matching requirements (see, e.g., Ref. 6).

For active emulation, a shaker must meet or exceed the force magnitude required within the frequency range of interest. The required magnitude is strongly influenced by the shaker's mounting location in the passive emulator. Consequently, while any location satisfying the passive matching requirements is possible, the emulator's base frame is often the best location since it typically maximizes the transfer accelerance to the emulator mounting locations permitting use of a smaller shaker.

Given a set of emulating shaker mounting locations, the forces which they should produce can be computed using $A_{fi}^{e}(\omega)$, the transfer accelerance matrix relating shaker forces to emulator mounting location accelerations. This accelerance matrix can be experimentally determined or theoretically obtained from the passive emulator model. The shaker forces $F_i^e(\omega)$ are the solution of

$$a_f^a(\omega) = A_{fi}^{e}(\omega) F_i^e(\omega). \tag{3}$$

Considering linear coordinates (for simplicity of presentation), an upper bound on the number of emulating shakers is given by $\dim[a_f^a(\omega)]$, i.e., one for every coordinate direction at every mounting location. In this case, $A_{fi}^{e}(\omega)$ is a square matrix and Eq. (3) is solved by matrix inversion over the frequency range of interest.

Since an emulator models a single machine or piece of equipment, in many practical situations, the actual number of emulating shakers needed is less than $\dim[a_f^a(\omega)]$. Mathematically, a surplus of shakers would appear in Eq. (3) as a loss of column rank in $A_{fi}^{e}(\omega)$. By examining $A_{fi}^{e}(\omega)$, one can deduce which emulating shakers can be omitted while still ensuring that $a_f^a(\omega)$ lies in the column space of $A_{fi}^{e}(\omega)$. In these cases, $\dim[F_i^e(\omega)] < \dim[a_f^a(\omega)]$ and a least squares solution of Eq. (3) can be employed.

2.2 Emulating shaker controller

For the shakers to produce the forces calculated from Eq. (3), they must be driven under closed-loop control. For example, the controller of Fig. 2 utilizes proportional and derivative (PD) feedback together with a feedforward term to obtain the voltage to be applied to the emulating shaker. The dynamic model, $V(s)/F(s)$, relating shaker voltage and force over the frequency range of interest, can be readily obtained from experiment using model estimation of sinusoidal sweep data.

In calculating the emulator forces $F_i^e(t)$ from Eq. (3), if the acceleration is composed of several dominant harmonics, the force is likely to be dominated by these frequencies. In this

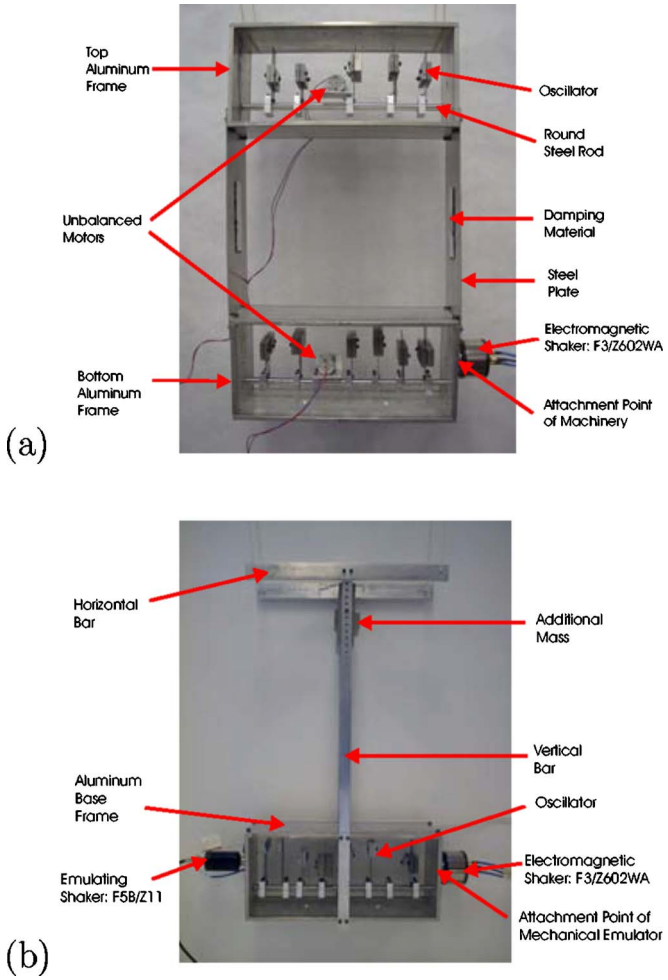


Fig. 3. (Color online) (a) Machinery test bed. (b) Emulator test bed. Both systems are suspended for testing.

case, the amplitude and phase angle of each frequency component in the forces can be evaluated discretely. Otherwise, the time-domain forces are obtained from the inverse Fourier transformation of $F_i^e(\omega)$.

3. Experimental example

To evaluate the proposed approach to active emulation, a machinery test bed was constructed as shown in Fig. 3(a). Use of this test bed instead of an actual machine provided the ability to vary the number, frequency, and damping of the passive modes as well as the internal forcing.

The depicted system is comprised of two aluminum frames connected by steel plates. A total of 11 oscillators are mounted on the top and bottom frames whose fixed-base frequencies can be tuned between 10 Hz and 80 Hz. The system also possesses a structural mode at about 25 Hz associated with bending of the steel plates. Polymeric damping materials are attached to both the oscillators and the frame to achieve various levels of damping. Two dc motors with eccentric masses fixed to their shafts are mounted as shown in the figure. These motors, when not rotating, contribute two modes between 60 Hz and 70 Hz. The height of the test bed is 56 cm, its width is 30 cm, and its depth is 10 cm. The total mass of the test bed is 6.12 kg.

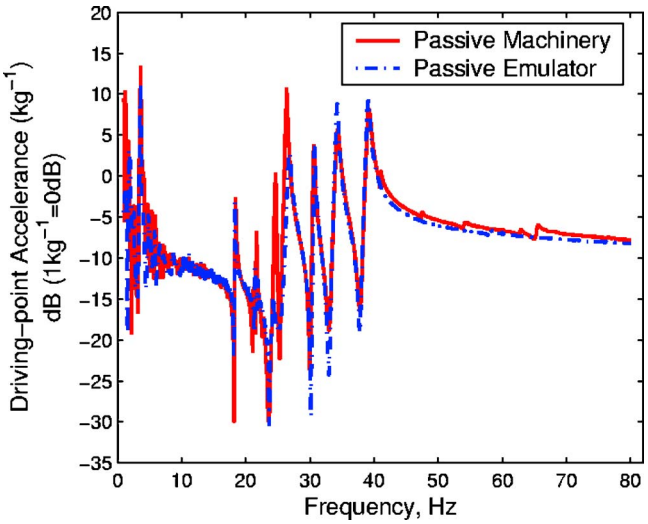


Fig. 4. (Color online) Horizontal mounting-point passive accelerance of machinery test bed and emulator.

A single superstructure mounting location is defined to be at the lower right corner where the testing shaker is shown attached. Three degrees of freedom are considered for this mounting location consisting of translations in the horizontal and vertical directions and rotations in the plane of the figure. As depicted, the test bed is suspended from four cords for testing purposes.

The passive emulator of Fig. 3(b) was designed using methods described in Ref. 6 to provide rigid-body matching in the vertical and rotational coordinate directions and dynamic matching between 10–80 Hz in the horizontal direction at the mounting location. A comparison of the horizontal direction accelerance is given in Fig. 4. Note that the two modes below 10 Hz are rigid-body suspension modes.

For active emulation, a single emulating shaker was needed. Its mounting location was selected on the base frame as shown in Fig. 3(b). The total mass of the emulating shaker and

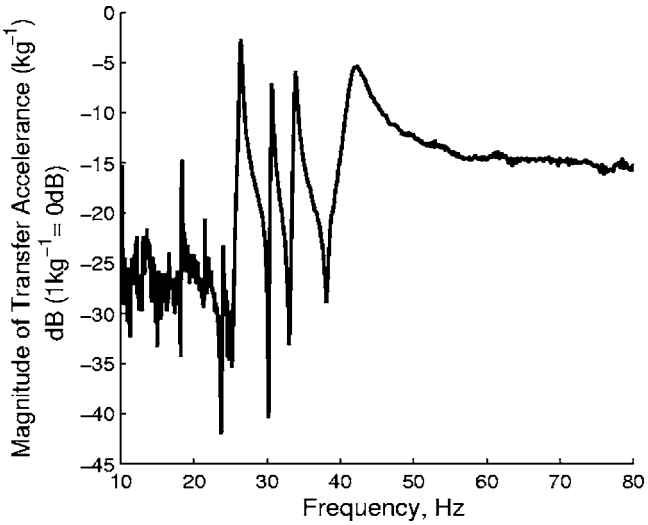


FIG. 5. Transfer accelerance, $A^e_{fi}(\omega)$, of emulating shaker.

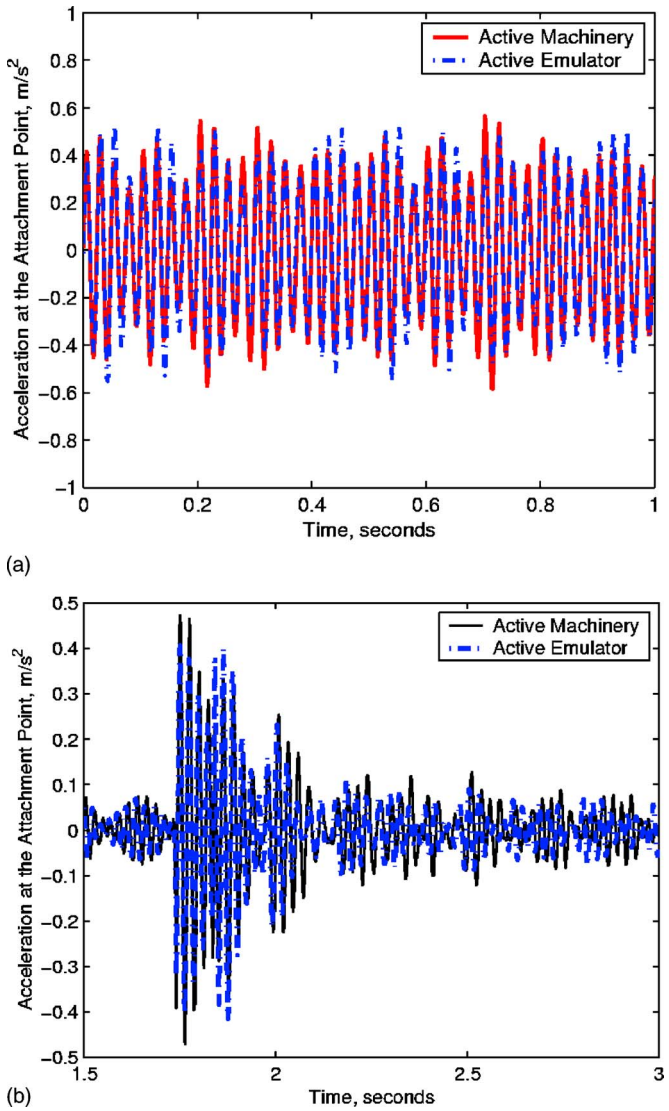


FIG. 6. (Color online) Response of active machinery and emulator. (a) Steady-state loading from superstructure. (b) Shock loading from superstructure.

impedance head is 0.170 kg. The transfer accelerance $A_{fi}^{e}(\omega)$ from the emulating shaker location to the superstructure mounting location was measured between 10–80 Hz, as shown in Fig. 5.

To implement the controller of Fig. 2, the transfer function, $V(s)/F(s)$, was estimated from sinusoidal sweep data between 10–80 Hz to be

$$\frac{V(s)}{F(s)} = -\frac{2.007s^2 + 226.5s + 86670}{s^2 + 20s + 455}. \tag{4}$$

The feedback controller was implemented using a 16-bit AD/DA board and a personal computer.

Two experiments were carried out to verify active emulation. In both tests, the unbalanced motors of the machinery test bed were driven at frequencies of approximately 38 and

44 Hz. Given the discrete nature of the forcing, the emulating shaker forces were calculated using Eq. (3) and the data of Fig. 5 evaluated at the two discrete frequencies.

The first experiment simulates the case in which the superstructure applies periodic forcing to the system, such as might be produced by nearby machinery. The second experiment simulates the situation of the superstructure applying a shock loading to the system. These forcing signals were applied using a shaker located at the superstructure mounting location and the resulting force and acceleration were recorded using an impedance head.

3.1 Experiment 1: Periodic excitation

In this experiment, the horizontal forcing applied to the machinery and emulator mounting points was a weighted sum of five sinusoids of frequencies 20, 28, 32, 40, and 50 Hz. The mounting location response of the machinery test bed and emulator are depicted in Fig. 6(a). This is the total response due to forcing by the superstructure and by the internal active loading. The relative acceleration error magnitudes at the five superstructure forcing frequencies are all less than 10% and can be attributed to passive emulation error. These relatively large errors can be anticipated since the frequencies lie in a region of Fig. 4 in which large changes in magnitude occur for small changes in frequency. At the active vibration frequencies of 37 and 44 Hz, the relative emulation error magnitudes are a more modest 6.67% and 1.13%, respectively. Note from the transfer accelerance of Fig. 5, however, that the active emulation error can be similarly dependent on frequency.

3.2 Experiment 2: Impulsive excitation

The results of the impulsive loading experiment are shown in Fig. 6(b). The impulse is applied at about 1.75 s. The signal before this time corresponds to the emulating shaker matching the mounting location acceleration produced by the two unbalanced machinery motors. The relative magnitude error at the two motor frequencies is 3.23% and 0.95%. Just after application of the impulse, the response is dominated by the passive dynamics of the systems which are in good agreement. As the responses decay, however, small differences in the passive dynamics can be seen for times 2.25–3 s. Later times, not depicted, show recovery to the level of matching depicted prior to 1.75 s.

4. Conclusions

A two-step design process for active electromechanical emulation has been developed. In the first step, a passive mechanical emulator is designed to match the passive dynamic properties of a vibratory system. In the next step, this passive emulator is modified by the addition of active elements driven by a controller to reproduce the mounting location accelerations of the active vibratory system. While the approach has been demonstrated on a planar test bed, the extension of the method to higher dimensions is straightforward.

Acknowledgments

This work was supported by the Office of Naval Research under Grants Nos. N00014-01-1-0155 and N00014-03-1-0881.

References and links

[1]G. M. L. Gladwell, "Inverse Problems in Vibration," Appl. Mech. Rev. **49**, S25–S34 (1996).
[2]S. L. Chen and M. Géradin, "An Exact Model Reduction Procedure for Mechanical Systems," Comput. Methods Appl. Mech. Eng. **143**, 69–78 (1997).
[3]A. D. Pierce, "Resonant-Frequency-Distribution of Internal Mass Inferred from Mechanical Impedance Matrices, with Application to Fuzzy Structure Theory," ASME J. Vibr. Acoust. **119**, 325–333 (1997).
[4]W. Chen and P. E. Dupont, "Realization of Mechanical Systems from Second-Order Models," J. Acoust. Soc. Am. **118**, 762–773 (2005).
[5]L. Kari, "On the Dynamic Stiffness of Preloaded Vibration Isolators in the Audible Frequency Range: Modeling and Experiments," J. Acoust. Soc. Am. **113**, 1909–1921 (2003).
[6]W. Chen, "Mechanical Realization Theory and Its Application to Machinery Emulation," Ph.D. dissertation, College of Engineering, Boston University, Boston, MA, 2004.

Classification of vocalizations of killer whales using dynamic time warping

Judith C. Brown
Physics Department, Wellesley College, Wellesley, Massachusetts 02481 and Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139
brown@media.mit.edu

Andrea Hodgins-Davis
Biology Department, Wellesley College, Wellesley, Massachusetts 02481
ahodgins-davis@mbl.edu

Patrick J. O. Miller
Sea Mammal Research Unit, University of St. Andrews, St. Andrews, Fife KY16 9QQ, United Kingdom
pm29@st-andrews.ac.uk

Abstract: A large number of killer whale sounds have recently been classified perceptually into Call Types. [A. Hodgins-Davis, thesis, Wellesley College (2004)]. The repetition rate of the pulsed component of five or more examples of each call type has been calculated using a modified form of the FFT based comb-filter method. A dissimilarity or distance matrix for these sounds was calculated using dynamic time warping to compare their melodic contours. These distances were transformed into a component space using multidimensional scaling and the resulting points were clustered with a kmeans algorithm. In grouping 57 sounds into 9 call types, a single discrepancy between the perceptual and the automated methods occurred.

PACS numbers: 43.80.Ka [CFM]
Date Received: October 10, 2005 **Date Accepted:** December 19, 2005

1. Introduction

Marine mammals produce a wide range of vocalizations, and an improved ability to classify recorded sounds could aid in species identification or in tracking movements of animal groups. In the case of killer whales, time-frequency contours of stereotyped pulsed calls are group-specific from matrilineal lines (group with same mother) to pods (living group consisting of a number of matrilineal lines) to clans (larger groups sharing calls) (Ford, 1991; Miller and Bain, 2000). There are a number of call types within these groups, which are thought to be learned in the pod, and for the most part, these types have been classified by humans from listening and looking at their spectra. For killer whale sounds classification by eye and ear is quite consistent (Yurk *et al.*, 2002), and this type of classification has been useful to reveal group-specific acoustic repertoires and matching vocal exchanges (Ford, 1991; Miller *et al.*, 2004).

It would, nonetheless, be useful to replace human classification with an automatic technique because of the large amounts of data to be classified, and the fact that automatic methods can be fully replicated in subsequent studies. In the last decade, there have been attempts at automatic classification, including a single dynamic time warping study (DTW). Buck and Tyack (1993) took three signature (specific to the individual) whistles from each of five dolphins and used DTW to compare to five reference whistles, one from each of the five dolphins. Correct identification was achieved for all 15 of the signature whistles.

Killer whales produce three forms of vocalizations: clicks, whistles, and pulsed calls. Clicks consist of an impulse train (series of broadband pulses); whistles consist, for the most part, of a single sinusoid with varying frequency; and pulsed calls are more complex sounds

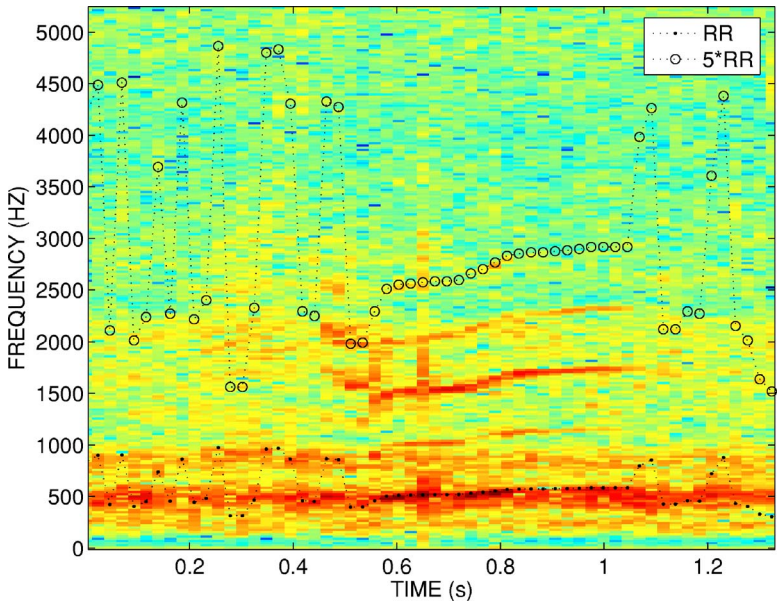


FIG. 1. (Color online) Spectrogram of a sound with fundamental frequency (RR for repetition rate) and fifth harmonic (5*RR) displayed. This sound has a fundamental frequency which falls in a strong band of noise. Noise alone can be seen before and after the call.

with many harmonics. The repetition rate is a measure of the periodicity of the signal, and its measurement is called "pitch tracking" or "fundamental frequency tracking" in the speech literature.

Our classification of pulsed whale sounds is a much more difficult problem than the previous work using whistles. There are many more calls here, and these pulsed calls are far more complex than whistles which generally consist of a single harmonic eliminating the difficult step of obtaining the repetition rate. No reference calls are used for our study; all sounds are clustered with no assumptions other than the number of types previously classified perceptually.

2. Repetition rate measurement

A large number of sounds from the Captive Killer Whale Population at Marineland of Antibes, France were recorded using an HTI hydrophone directly to a computer hard-drive. These were recently classified into call types by listening to the calls and examining their spectrograms (Hodgins-Davis, 2004). Five or more examples of each of nine perceptually identified call types of these killer whale vocalizations were chosen for this study. These sounds were among those with the highest signal to noise ratio estimated visually from their spectrograms.

Repetition rates based on calculation of the autocorrelation function (Brown and Zhang, 1991) and the constant Q transform (Brown, 1992) were measured initially since these have proven accurate for fundamental frequency tracking of musical sounds. A third method (Brown *et al.*, 2004), called the comb filter method, proved more effective in dealing with the principal difficulty, which is the noise of the water circulation pump in some frequency bands. The comb-filter method is FFT-based and involves adding up a variable number of Fourier components with a fixed spacing (Hess, 1983); it can be adapted to accommodate several input parameters:

(1) The upper and lower range of frequencies which are searched as possible repetition rate candidates.

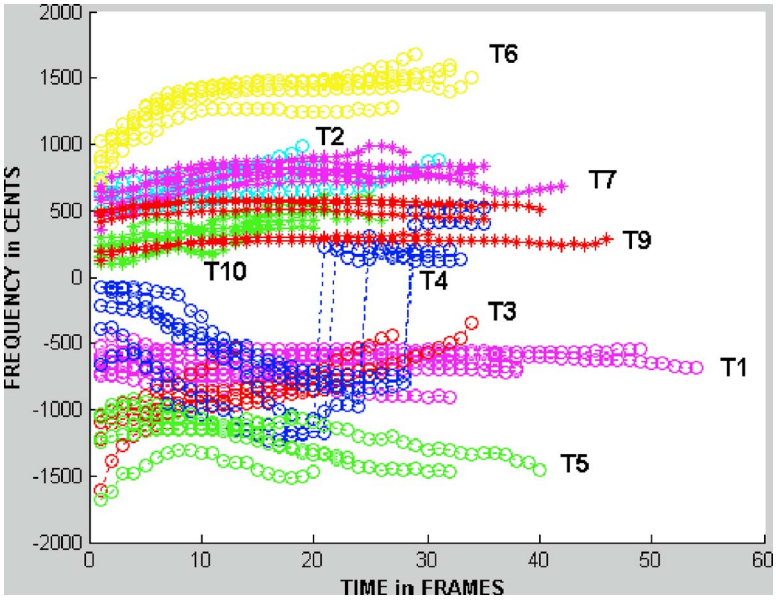


FIG. 2. (Color online) Fundamental frequency contours in cents for each of the sounds with perceptual grouping indicated. The total time for 60 frames is 1.4 s. Frequencies in Hz range roughly from 185 Hz (−1500 cents) to 1050 Hz (1500 cents). This figure is much clearer in color in the online publication.

(2) The lowest frequency at which the comb is fit. This made it possible to jump over the pump noise which in general was below around 500 Hz.
(3) The number of "teeth" (harmonics) in the comb.

The FFT was calculated with a frame size of 46.4 ms (window size of 1024 at sample rate 22050) and hop size of 23.2 ms giving 50% overlap and a frequency resolution of 21.6 Hz. The FFT was upsampled by a factor of 10 prior to applying the "comb" allowing identification of peak position to within 2.2 Hz. Parameters were adjusted for each of the call types, and the calls were extracted manually.

An example of these results is found in Fig. 1 with audio in Mm. 1. The noise in this figure before the call starts can be easily identified. The pitch track is graphed for both the fundamental and a higher harmonic, in this case the fifth. This was necessary for some of the calls with a fundamental frequency buried in the noise, such as this one, where only the higher harmonics could be seen clearly. It is also a much better visual indicator of errors for a low frequency fundamental where a small error in the fundamental is difficult to see, but is multiplied by the harmonic number for the higher harmonic and becomes more evident.

Mm. 1 Audio file in wav format (60 kb) for the example shown in Fig. 1.

A summary figure including all fundamental frequency contours for all types is given in Fig. 2. Frequencies are measured in cents defined as

$$f_{\text{cents}} = 1200 \cdot \log_2(f/f_{\text{ref}}), \quad (1)$$

where we have chosen the reference frequency f_{ref} as 440 Hz. This unit is commonly used for musical frequencies, and gives a 100 cent increase for each semitone of the equal tempered scale which is roughly 6% of the frequency. This is a geometric measure in that equal percentage changes are transformed into equal additive differences. The advantage is that a given melody has the same shape independent of its initial frequency. For example, consider doubling the frequency from 100 Hz to 200 Hz compared to 400 Hz to 800 Hz. In each of these cases the change measured in cents is 1200.

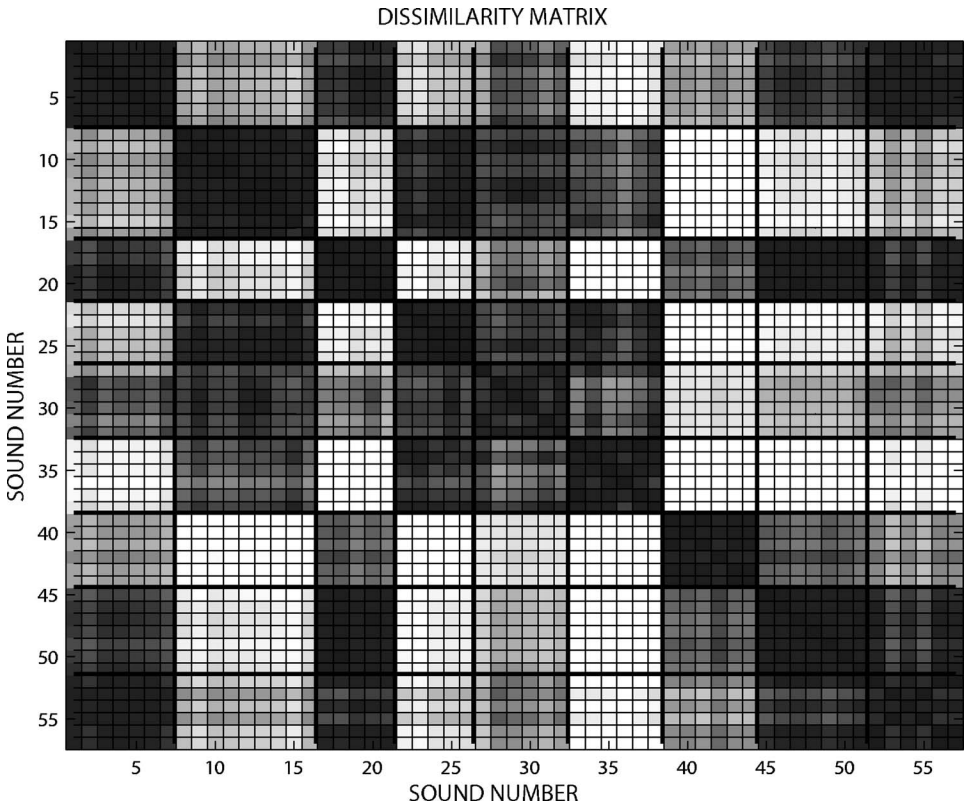


FIG. 3. Dissimilarity matrix calculated with dynamic time warping for fundamental frequency contours of Fig. 2. The sound blocks are in the order T10, T1, T2, T3, T4, T5, T6, T7, T9 as labeled in Fig. 2.

In Fig. 2 some call types are well-separated in frequency, while others overlap considerably. This figure distinguishes between the types much more clearly in color in the online publication than in black and white.

3. Dynamic time warping

The pitch contours were compared pairwise using the dynamic time warping (DTW) method described in Chai and Vercoe (2003). See also Rabiner and Juang (1993) for a thorough discussion of methods of dynamic time warping. The frequency values in cents from Eq. (1) were divided by 100 to retain the logarithmic measure of frequency but to have one or two digit differences. This facilitated a quick grasp of the numerical values involved for choosing parameter values.

The fundamental frequencies of all possible pairs of sounds were compared number by number. We will refer to a typical pair as sound 1 and sound 2. A scoring matrix was constructed using the algorithm (Chai and Vercoe, 2003):

$$M[1,1] = 0,$$

$$M[1,j] = M[1,j-1] + b,$$

$$M[i,1] = M[i-1,1] + a,$$

and

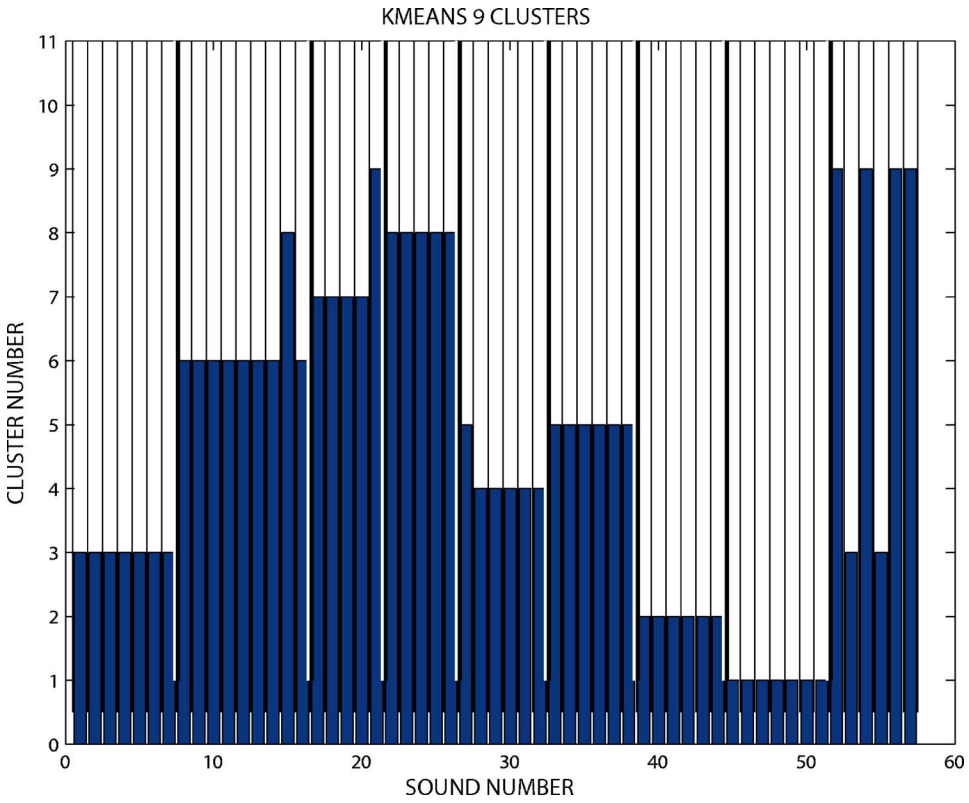


FIG. 4. (Color online) Clustering results for dissimilarity matrix for differences in fundamental frequency contours alone.

$$M[i,j] = \min(M[i-1,j] + a, M[i,j-1] + b, M[i-1,j-1] + D), \tag{2}$$

where $i,j \geqslant 2$ and D is the difference in the jth frequency of sound 1 and the ith frequency of sound 2. Here a is the cost of an insertion and b is the cost of deletion.

The resulting score is the lowest value in the last row normalized by dividing by the length of the shorter sound. The matrix of scores with one score tabulated for each pair of sounds is shown in the dissimilarity (or distance) matrix of Fig. 3. For identification purposes, blocks which were grouped in the previous perceptual classification are outlined by bold lines. Perfect agreement with the perceptual classification, would give dark (low number for small distance) in the blocks along the diagonal (comparisons within the same perceptual call type), and white elsewhere, corresponding to large differences between dissimilar call types. Dark blocks off the diagonal indicate call types which are similar and difficult to separate in classification.

4. Results

4.1 Experiment 1

Adjustable parameters in Eq. (2) for the calculation of the dissimilarity matrix of Fig. 3 are the costs a and b of insertions and deletions. The deletion cost corresponds to a difference in the duration of a call rather than the frequency contour so b was chosen to be zero. An initial guess of 15 was made for the insertion cost, and this turned out to be very near optimum for matching the perceptual results. This value was twice the average deviation from the mean of the numbers for the melodic contours.

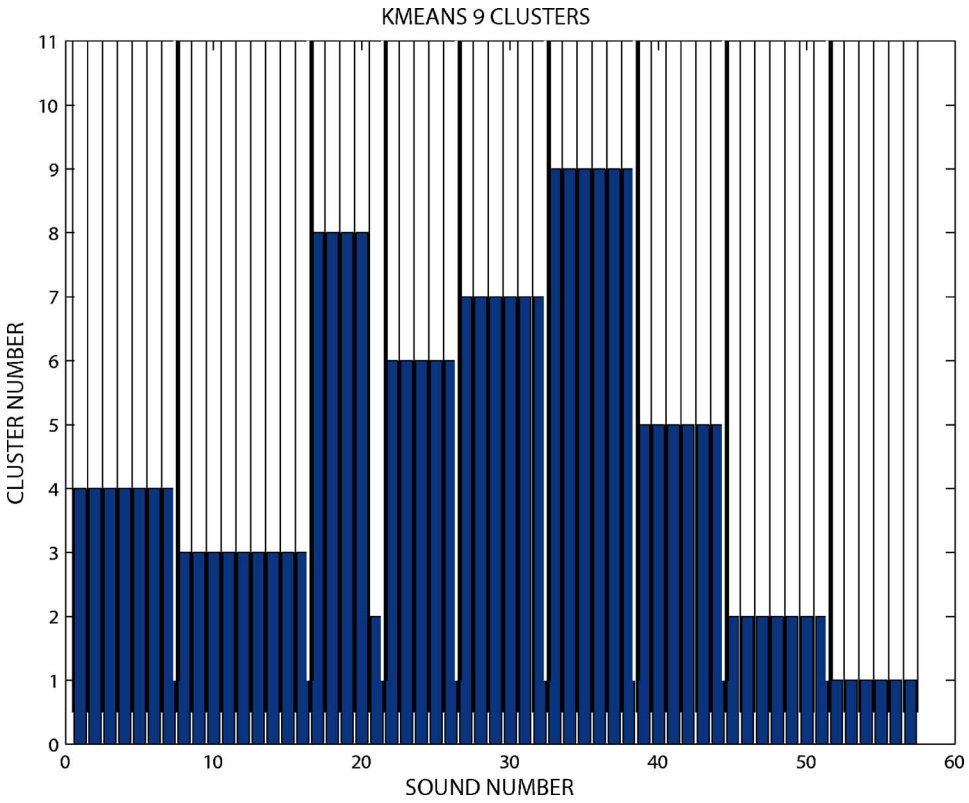


FIG. 5. (Color online) Clustering results for sum of dissimilarity matrices for differences in fundamental frequency contours and for derivatives (point to point differences) of the frequency contours shown in Fig. 3.

To obtain clusters from the distance values in the dissimilarity matrix, we first used multidimensional scaling (MDS) with matlab function *cmdscale*. MDS is a tool usually used for dimensionality reduction and/or data visualization (Kruskal and Wish, 1978). For example one can take perceptual differences in sounds such as those generated by musical instruments, transform them with MDS into points in a space with as few dimensions as possible, and then attempt to associate those dimensions with acoustical correlates. The mathematics of this transformation from distances to a coordinate representation is presented clearly in Borg and Groenen (1997).

For our purposes MDS was used to transform the dissimilarity values into positions in a Euclidian-type space; they could then be grouped using a clustering algorithm, implemented in this case with matlab function *kmeans*.

The number of output clusters was chosen to be 9 corresponding to the number of perceptual clusters. Clustering algorithms are nondeterministic as they depend on random initial conditions, and risk finding local rather than global minima. To insure that the global (true) minimum is found, a large number of runs must be made. Out of 100 total runs, in over 90 of these, the number of errors (compared to perceptual results) was 5 or 7 out of a total of 57 sounds. A typical classification can be found in Fig. 4, which is a bar graph of the cluster assignment vs sound numbers. Any of the results with 5 errors would be equivalent to this one but would have different cluster assignments on the vertical axis. These numbers are not significant except in showing that different call types are separated into different clusters.

The vertical boldface black lines show the breaks in perceptually defined types. For example the first perceptual group consists of sounds 1–7 and these are assigned to the same cluster with no errors; this means that all sounds judged by a human to belong to the same call

type are classified by DTW differences into the same cluster. There is one error each in the second, third, and fifth perceptual groups. The last perceptual group has two of its members assigned to the first group for two errors; so in this example there are five errors overall for an 88% agreement with perceptual results.

The DTW parameter a was varied to determine its effect on the results. If it is roughly halved from 15 to 8, the errors increase slightly to 8, but also there is mixing of types 2 and 7. From Fig. 2 it can be seen that these curves are very similar.

4.2 Experiment 2

The DTW distances as calculated for Fig. 3 weigh the similarity of absolute frequencies more than the shape of the contour. The derivative of these curves, calculated as point to point differences of the contour, is a measure of the shape alone as the absolute frequency is subtracted out. To combine these effects a dissimilarity matrix for the derivatives was calculated using the same procedure as described for the contours. Since these numbers for the differences are roughly a factor of 10 (ratio of their standard deviations is 9.8) smaller than those for the absolute frequencies, this matrix was multiplied by 10 and added to the matrix of Fig. 3. This sum takes both absolute frequencies and contours into account. Multidimensional scaling and kmeans calculations were performed as before. In over 80% of the runs, the results shown in Fig. 5 give a single sound differing from the perceptual classification for essentially perfect agreement.

5. Conclusions

Dynamic time warping has proven very successful for the automatic classification of killer whale vocalizations with the limitation that the determination of the melodic contours of the input sounds can be time-consuming. Further testing with diverse call repertoires recorded under natural conditions is planned in order to determine the full potential of this technique.

Acknowledgements

J.C.B. is very grateful to Wei Chai, who provided her matlab code for DTW, plus much helpful advice on its use and on dynamic programming in general. Thanks to Marineland, Antibes for support installing and using the pool hydrophones. Funding was provided by WHOI's Ocean Life Institute and a Royal Society fellowship to P.J.O.M.

References and links

Borg, I., and Groenen, P. (**1997**). *Modern Multidimensional Scaling. Theory and Applications*, Springer, New York.

Brown, J. C., and Zhang, B. (**1991**). "Musical frequency tracking using the methods of conventional and 'narrowed' autocorrelation," J. Acoust. Soc. Am. **89**, 2346–2354.

Brown, J. C. (**1992**). "Musical fundamental frequency tracking using a pattern recognition method," J. Acoust. Soc. Am. **92**, 1394–1402.

Brown, J. C., Hodgins-Davis, A., and Miller, P. J. O. (**2004**). "Calculation of repetition rates of the vocalizations of killer whales," J. Acoust. Soc. Am. **116**, 2615.

Buck, J. R., and Tyack, P. L. (**1993**). "A quantitative measure of similarity for Tursiops truncatus signature whistles," J. Acoust. Soc. Am. **94**, 2497–2506.

Chai, Wei, and Vercoe, Barry (**2003**). "Structural analysis of musical signals for indexing and thumbnailing," in Proceedings of ACM/IEEE Joint Conference on Digital Libraries.

Ford, J. K. B. (**1991**). "Vocal traditions among resident killer whales Orcinus orca in coastal waters of British Columbia," Can. J. Zool. **69**, 1454–1483.

Hess, W. (**1983**). *Pitch Determination of Speech Signals: Algorithms and Devices*, Springer-Verlag, Berlin.

Hodgins-Davis, A. (**2004**). "An analysis of the vocal repertoire of the captive killer whale population at Marineland of Antibes, France," thesis, Wellesley College.

Kruskal, J. B., and Wish, M. (**1978**). *Multidimensional Scaling*, Sage Publications, Beverly Hills, CA.

Miller, P. J. O. and Bain, D. E. (**2000**). "Within-pod variation in the sound production of a pod of killer whales, Orcinus orca," Anim. Behav. **60**, 617–628.

Miller, P. J. O., Shapiro, A. D., Tyack, P. L., and Solow, A. R. (**2004**). "Call-type matching in vocal exchanges of free-ranging resident killer whales, Orcinus orca," Anim. Behav. **67**, 1099–1107.

Rabiner, L., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition*, Prentice-Hall, New Jersey.

Yurk, H., Barrett-Lennard, L., Ford, J. K. B., and Matkin, C. O. (**2002**). "Cultural transmission within maternal lineages: vocal clans in resident killer whales in southern Alaska," Anim. Behav. **63**, 1103–1119.

Auditory brainstem response in a harbor porpoise show lack of automatic gain control for simulated echoes

Kristian Beedholm and Lee A. Miller
Institute of Biology, University of Southern Denmark, Campusvej 55, DK 5230 Odense M, Denmark
beedholm@mail.dk, lee@biology.sdu.dk

Marie-Anne Blanchet
Fjord & Bælt, Margrethes Plads 1, DK 5300 Kerteminde, Denmark
marie@fjord-baelt.dk

Abstract: The auditory brainstem response (ABR) response to simulated echolocation clicks was studied in a harbor porpoise, *Phocoena phocoena*, to determine the relationship between the animal's perceived echo strength and the simulated target distance. In one experiment the click level at the listening post was kept constant while delay was changed, in another, the level was varied to approximate spreading losses. Results of both experiments indicate that there is no automatic gain control in the hearing system of this harbor porpoise.

PACS numbers: 43.80.Lb, 43.64.Ri [CFM]
Date Received: November 9, 2005 **Date Accepted:** December 19, 2005

1. Introduction

The auditory brainstem response (ABR) technique is being used more often for studying hearing in toothed whales (odontocetes) mainly because it is noninvasive and the suction cup EEG electrodes are easy to apply. With this method an acoustic stimulus is presented to the animal several hundreds of times. The collective electrical response (ABR) of the brain is time locked to the stimulus, recorded and averaged. The ABR response is particularly pronounced when elicited by short, broadband stimuli, such as a dolphin echolocation click. The brief click presumably activates many independent auditory channels.[1] The method is therefore suited for the study of echolocation. One study relevant to the results reported here showed that the amplitude of the ABR measured in a false killer whale, *Pseudorca crassidens*, remained unchanged when elicited by echoes from a physical target at different distances even though the level of the outgoing click was constant.[2,3] This result seems to indicate that this odontocete might possess some form of automatic gain control (AGC) in its hearing system, as proposed for echolocating bats.[4–7] Such a mechanism regulates the amplitude of the perceived echo over range to the target. In bats both peripheral and central mechanisms provide automatic gain control.[8,9] Contraction of middle ear muscles during the intense vocalization protects the inner ear and the gradual relaxation of this protective mechanism contributes to the increased sensitivity for targets at longer distances. If a similar mechanism were present in odontocetes it would have to operate at a much faster time scale due to the higher speed of sound in water and to the much shorter signals used by odontocetes. Little is known about middle ear function in odontocetes.[10]

In this study we addressed the question of a possible AGC mechanism in the auditory system of a harbor porpoise. We used a target simulator, which allows for easy uncoupling of the simulated echo amplitude and the simulated target distance. Two experiments were run. In the first the amplitude of the simulated echo was kept constant regardless of the delay. In the second experiment the amplitude of the simulated echo was adjusted slightly less than would be predicted for two-way spherical spreading losses. We found no evidence for an AGC mechanism in a trained harbor porpoise.

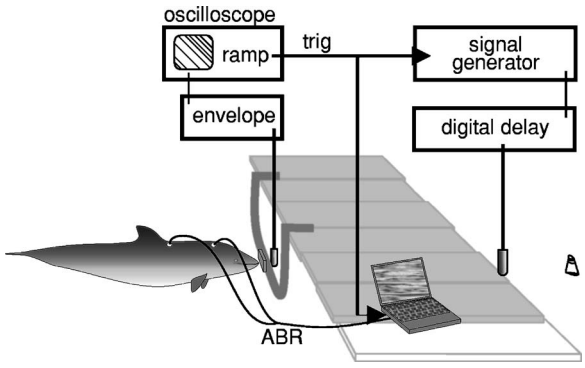


Fig. 1. Schematic of the set-up and some of the equipment used to produce simulated echoes and record ABRs. The small ghost to the right of the transmitting hydrophone symbolizes the simulated target at one possible virtual location. The schematic is not to scale.

2. Methods

We used a 9-years old male harbor porpoise, Eigil, who was born in the wild but lived most of his life in captivity. The experiments were carried out at Fjord and Bælt in Kerteminde, Denmark. Eigil was trained to station 1 m below the water surface with his rostrum on a small plastic square (the listening post) where he could station there for up to 90 s (see Mm. 1.). Before going to the listening post he was fitted with silver electrodes in suction cups, one behind the blowhole and the second near the dorsal fin, the reference electrode. He always vigorously echolocated during trials producing clicks at rates up to 300 Hz. There were normally four to eight trials per session and one or two sessions per day.

The simulated echo was the sampled step response from a 1/3-octave filter set at 125 kHz. This signal is very similar to a porpoise click in time frequency structure due to the fact that both are minimum phase bandpass signals.[11] The simulated echo was stored in and delivered by an Agilent arbitrary waveform generator (33220A). The porpoise's echolocation clicks were transduced by a Reson TC4013 hydrophone placed 0.5 m in front of the listening post. The envelopes of clicks were obtained from a click detector and used to trigger the simulated echo. The threshold for triggering was set to 132 dB re. 1 μPa peRMS (peak equivalent root mean square[12]) at the hydrophone. A digital delay channel was used to generate the desired delays. The amplitude was set with a digitally controlled attenuator. The amplified simulated echo was projected from a Sonar Products HS150 hydrophone at the backside of the pontoon bridge, 2.3 m in front of the listening post (Fig. 1).

The trigger signal for the simulated echo from the generator also served to trigger a 16 or 32 ms recording period containing the ABR, the potential difference between the two electrodes on the porpoise amplified by 110 dB. The recording period containing the ABR was sampled at 250 or 500 ksamples/s with a 12 bit dynamic range (National Instruments, DAQCard-6062E), averaged on line and saved after 64 presentations. The extreme over sampling eliminated the risk of high frequency interference from diverse sources showing up as aliased components within the frequency band of interest. The averaged traces were digitally filtered off-line between 500–2200 Hz, the frequency band where the ABR signal had the most energy, and the peak of the absolute amplitude was registered.

In experiment 1, the amplitude of the simulated echo at the listening post was held constant at 128 dB re. 1 μPa peRMS while the delay between the porpoise click and simulated echo was changed, representing changing distances to the simulated target. In experiment 2, the amplitude of the simulated echo was reduced by 9 dB per doubling of delay or somewhat less than expected from spherical spreading losses; −12 dB per doubling of distance. A perfect auditory AGC mechanism should compensate with +12 dB per doubling of distance, so by using 9 dB, the amplitude of the ABR response should *increase* with increasing delay. The peak ABR

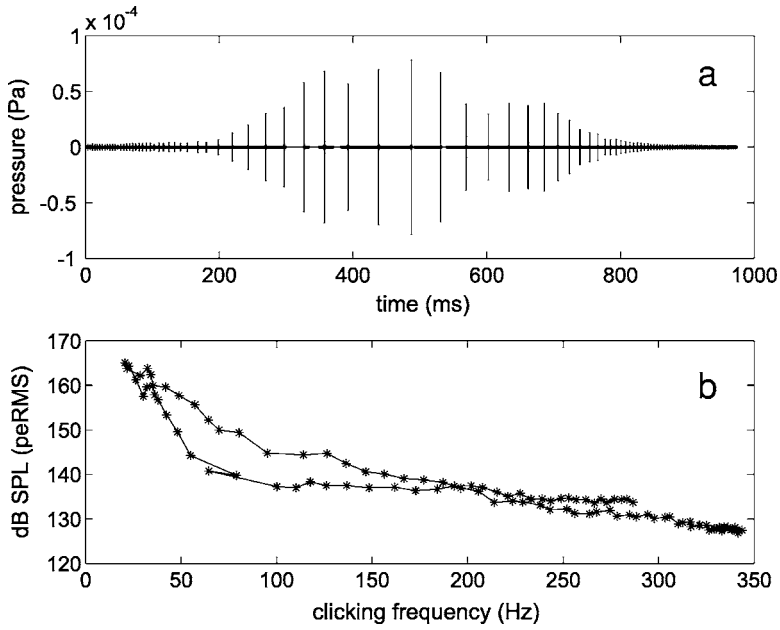


Fig. 2. Example of Eigil's echolocation behavior while on station. (**a**) Raw sample data recorded in slightly less than one second while Eigil focused his attention in the direction of the simulated target (documented by simultaneous video recordings-Mm. 1.). Click amplitudes are clearly negatively correlated with click rate. In (**b**) this relationship is documented further by plotting peak amplitude against the click rate defined as the reciprocal of the time to the preceding click.

levels in experiment 2 were compared to the peak ABR levels recorded at different amplitudes, but at a constant delay of 8.5 ms, which is the average of all delays tested in experiment 2. Two sessions on the same day with 5 trials each were used to gather data for experiment 1 and one session on another day with 5 trials was used for data gathering in experiment 2.

3. Results

Figure 2 shows part of a typical echolocation series while Eigil is on station. He spontaneously alternates between high intensity, low rate clicks and low intensity, high rate clicks several times during a trial. Only high intensity signals triggered simulated echoes. He is rewarded for remaining on station and not for producing sonar clicks.

[Mm. 1. Eigil at station, sonar clicks via a detector (562 KB).]

The results of experiments 1 and 2 are shown in the form of waterfall plots [Figs. 3(a) and 3(b)]. Each trace is displaced upwards by amounts proportional to the delay of the phantom target. In experiment 1 (constant echo levels) ABR peak amplitudes [Fig. 3(a)] show no obvious changes as the delay increased. Also, the ABR elicited by the echolocation pulse is constant in amplitude [first ABR in each trace in Fig. 3(a)]. In experiment 2 (decreasing echo levels with increased delay) ABR peak amplitudes [Fig. 3(b)] decrease with increasing delay, where as the amplitudes of the ABR elicited by the outgoing clicks remain unchanged. The ABR amplitudes to the emission in experiment 2 [Fig. 3(b)] are larger at all stimulus delays than those in experiment 1 [Fig. 3(a)]. This may reflect a higher average outgoing click level in these trials, or it may reflect day-to-day variations in ABR amplitude. Figure 4 compares peak ABR amplitudes at various stimulus levels with a constant delay of 8.5 ms, the average delay in experiment 2, to ABR amplitudes obtained from experiment 2 [Fig. 3(b)]. The slope of 17.8 nV/dB, obtained

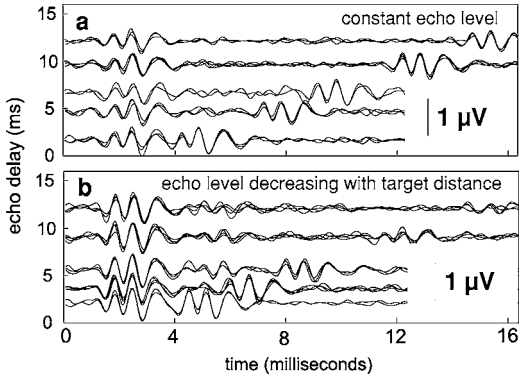


Fig. 3. ABR waveforms as a function of simulated echo delay. The *y*-axis indicates the delay for the simulated echo (2, 3.5, 5.5, 9, 12 ms). The vertical bar in the lower right-hand corner refers to ABR traces. Each trace is the average bioelectrical response to 64 stimulations. The first ABR deflection in each trace results from the outgoing porpoise click. In (**a**) the received level was held constant at 128 dB re. 1 μPa peRMS. Data were gathered from 5 trials in 2 sessions on one day. In (**b**) the echo stimulus levels were attenuated by 9 dB for a doubling of stimulus delay, approximating spherical spreading losses. Data were collected from 5 trials in 1 session.

when amplitudes decline with increased delay [see Fig. 3(b)], is not significantly different ($P < 0.57$) from the slope of 16.3 nV/dB obtained when the stimulus intensity was varied, but stimulus delay kept constant.

4. Conclusions

These results strongly suggest that this harbor porpoise possesses no automatic gain control (AGC), at least at the levels of the middle ear or the brainstem.

Had there been an active AGC mechanism in the first experiment, the response to the echo should have increased in amplitude with increasing delay, since a compensation for spreading losses would have allowed more sound energy to enter the hearing system with increasing delay. This did not happen [Fig. 3(a)].

In the second experiment where −9 dB was used for a doubling of delay, a perfect AGC mechanism in the sonar receiver would have given slightly increasing ABRs for increasing delays. This did not happen [Fig. 3(b)]. An incomplete AGC mechanism would have resulted in a regression coefficient between received echo level and ABR response that was lower than what we found when the echo delay was held constant, but echo amplitude varied (Fig. 4, filled circles). However, these slopes were not significantly different.

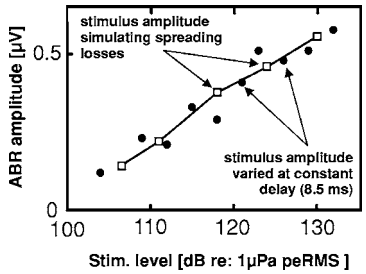


Fig. 4. Influence of echo level on peak ABR amplitude with and without accompanying delay changes. The peaks of the average data in Fig. 3(b) (one average value for each delay) are plotted against the received level in dB re. 1 μPa, peRMS (open squares, □). For comparison, the closed circles ● show the peak ABR values at various stimulus levels, but with a constant delay of 8.5 ms, the average delay for the results shown in Fig. 3(b).

One explanation for the marked difference to the results of Supin *et al.*[2,3] would be that the false killer whale adjusted the transmitted level as to keep the received echo at a constant amplitude and thus saturating the ABR to the outgoing click, obscuring the amplitude increase. But this scenario is unlikely, since the test animal produced very similar sounds at all delays.[3] In this respect the false killer whale's echolocation behavior differs from that recorded from free-ranging dolphins, where the transmitted click level is adjusted to the delay of the target by 6 dB per doubling of distance.[13,14]. We did not address the question of an AGC mechanism achieved by controlling the outgoing sonar click level since Eigil varies the click level almost continuously while on station (Fig. 2).

Another explanation might involve attention. In our experiments there was no reward associated with the echo, as was the case in the experiments of Supin *et al.*,[2,3] where the false killer whale performed a detection experiment.

Yet another possibility is that there might be a difference between the two species of odontocetes in this respect. If this were in fact the case, it might be explained by differences in source levels used by the two species. We never observed click levels above 180 dB re. 1 μPa peRMS from Eigil. It therefore might be argued that protection of the inner ear through some kind of middle ear mechanism is less likely in the smaller species than in the larger, louder toothed whales.

These questions could be resolvable if (1) the simulated echo experiments were performed with the false killer whale and/or (2) the harbor porpoise was trained to perform in a detection experiment[3] while the ABR recordings were made. We are currently planning the latter of these experiments.

Acknowledgements

This study was funded by the Office of Naval Research Award No. N000140210545. The harbor porpoises are maintained by Fjord & Bælt, Kerteminde, Denmark, under Permit No. J.nr. SN 343/FY-0014 and 1996-3446-0021 from the Danish Forest and Nature Agency, Danish Ministry of Environment. We acknowledge the staff at the Fjord and Bælt for their cooperation. We thank Paul Nachtigall for valuable discussions, and two reviewers for helpful and constructive remarks.

References and links

[1] A. Y. Supin, V. V. Popov, and A. M. Mass, *The Sensory Physiology of Aquatic Mammals* (Kluwer Academic, London, 2001).
[2] A. Y. Supin, P. E. Nachtigall, W. W. L. Au, and M. Breese, "The interaction of outgoing echolocation pulses and echoes in the false killer whale's auditory system: Evoked-potential study," J. Acoust. Soc. Am. **115**, 3218–3225 (2004).
[3] A. Y. Supin, P. E. Nachtigall, W. W. L. Au, and M. Breese, "Invariance of evoked-potential echo-responses to target strength and distance in an echolocating false killer whale," J. Acoust. Soc. Am. **117**, 3928–3935 (2005).
[4] S. A. Kick, and J. A. Simmons, "Automatic gain control in the bats sonar receiver and the neuroethology of echolocation," J. Neurosci. **4**, 2725–2737 (1984).
[5] J. A. Simmons, A. J. Moffat, and W. M. Masters, "Sonar gain control and echo detection thresholds in the echolocating bat, *Eptesicus fuscus*," J. Acoust. Soc. Am. **91**, 1150–1168 (1992).
[6] D. J. Hartley, "Stabilization of perceived echo amplitudes in echolocating bats. I. Echo detection and automatic gain control in the big brown bat, *Eptesicus fuscus*, and the fishing bat, *Noctilio leporinus*," J. Acoust. Soc. Am. **91**, 1120–1132 (1992a).
[7] D. J. Hartley, "Stabilization of perceived echo amplitudes in echolocating bats. II. The acoustic behavior of the big brown bat, *Eptesicus fuscus*, when tracking moving prey," J. Acoust. Soc. Am. **91**, 1133–1149 (1992b).
[8] O. W. Henson, "The activity and function of the middle-ear muscles in echo-locating bats," J. Physiol. (London) **180**, 871–887 (1965).
[9] N. Suga, and P. Schlegel, "Neural attenuation of responses to emitted sounds in echolocating bats," Science **177**, 82–84 (1972).
[10] S. H. Ridgway, D. A. Carder, T. Kamolnick, R. R. Smith, C. E. Schlundt, and W. R. Elsberry, "Hearing and whistling in the deep sea: depth influences whistle spectra but does not attenuate hearing by white whales (*Delphinapterus leucas*) (Odontocete, Cetacea)," J. Exp. Biol. **204**, 3829–3841 (2001).
[11] M. P. Olivieri, "What can be learned from one of nature's most advanced biosonars: Discussion on Bottlenose dolphins echolocation waveforms with respect to echolocation tasks in shallow water," J. Acoust. Soc. Am. **111**, 2371 (2002).

[12]D. R. Stapells, T. W. Picton, and A. D. Smith, "Normal hearing thresholds for clicks," J. Acoust. Soc. Am. **72**, 74–79 (1982).
[13]M. H. Rasmussen, L. A. Miller, and W. W. L. Au, "Source levels of clicks from free-ranging white-beaked dolphins (*Lagenorhynchus albirostris* Gray 1846) recorded in Icelandic waters," J. Acoust. Soc. Am. **111**, 1122–1125 (2002).
[14]W. W. L. Au and K. J. Benoit-Bird, "Automatic gain control in the echolocation system of dolphins," Nature (London) **423**, 861–863 (2003).

A hail size distribution impact transducer

John E. Lane
ASRC Aerospace Corporation, P.O. Box 21087, Kennedy Space Center, Florida 32815
John.Lane-1@ksc.nasa.gov

Robert C. Youngquist
National Aeronautics and Space Administration (NASA), Kennedy Space Center, Florida 32899
Robert.C.Youngquist@nasa.gov

William D. Haskell and Robert B. Cox
ASRC Aerospace Corporation, P.O. Box 21087, Kennedy Space Center, Florida 32815
William.Haskell-1@ksc.nasa.gov
Robert.Cox-1@ksc.nasa.gov

Abstract: An active impact transducer has been designed and tested for the purpose of monitoring hail fall in the vicinity of the Space Shuttle launch pads. An important outcome of this design is the opportunity to utilize frequency analysis to discriminate between the audio signal generated from raindrop impacts and that of hailstone impacts. The sound of hail impacting a metal plate is subtly but distinctly different from the sound of rain impacts. This useful characteristic permits application of signal processing algorithms that are inherently more robust than techniques relying on amplitude processing alone in the implementation of a hail disdrometer.

PACS numbers: 43.38.Yn, 43.60.Qv, 43.40.Dx, 43.28.Vd [AE]
Date Received: November 23, 2005 **Date Accepted:** December 16, 2005

1. Introduction

Impact disdrometers have long been a useful meteorological tool to measure and quantify rainfall drop size distributions, where a drop momentum is converted into a single electrical impulse.[1,2] The electrical impulse amplitude is converted to an estimate of drop diameter by means of an empirical calibration formula. The calibration may be a one-time procedure involving dropping numerous known-size-calibration drops from a tower of sufficient height to achieve terminal velocity. The test drop sizes must adequately span the size range of interest. Alternatively, or as a supplement to the single drop calibration, an *in situ* method of comparing the sum of accumulated drop impulses to the tip time interval of a tipping-bucket rain gauge may be utilized.[3]

Prior to launch, the Space Shuttle and other NASA launch vehicles are primarily large thermos bottles containing substantial quantities of cryogenic fuels. Because thermal insulation is a critical design requirement, the external wall of the launch vehicle fuel tank is covered with an insulating foam layer. This foam is fragile and can be damaged by very minor impacts, such as that from small- to medium-size hail, which may go unnoticed. In May 1999, hail damage to the top of the External Tank (ET) of STS-96 required a rollback from the launch pad to the Vehicle Assembly Building (VAB) for repair of the insulating foam. Because of the potential for hail damage to the ET while exposed to the weather, a vigilant hail sentry system using impact transducers can be deployed around the launch pads as a permanent warning system to record and quantify hail events, and to initiate a thorough inspection of the ET.

Since hail distributions are far less dense than raindrop distributions,[4] the surface area of a hail sensor must be much larger than that of a rainfall sensor. Overly large surface area is a detriment since the resulting overlap in drop impulses would make the instrument useless during intense storms. If the sensor collection area is too small, the measurements do not represent a good statistical sampling of the hydrometeor size distributions. Since the electrical impulse amplitude is well correlated to hydrometeor size, it is straightforward to measure and estimate

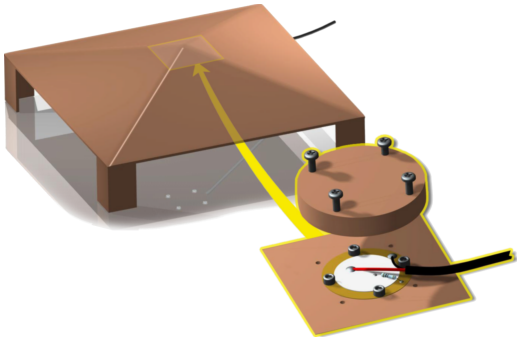


Fig. 1. (Color online) NASA/KSC hail transducer. The piezoelectric ceramic disk is shown in white and is mounted on a brass disk substrate.

hail size. However, small hailstones may be smaller than large raindrops. Therefore, impulse amplitude alone is insufficient for discriminating hail from rain in that size overlap region, a hail size regime common to the Florida climate.

A simple yet effective passive hail transducer is made using a 1-ft.,[2] 1-in.-thick Styrofoam panel, covered with thick aluminum foil.[5,6] This design is used extensively by the Colorado State University Department of Atmospheric Sciences and by over 3000 volunteers of the Community Collaborative Rain and Hail Study (CoCoRaHS) to measure and monitor hail in Colorado and adjacent states.[7] The Kennedy Space Center (KSC) hail transducer shown in Fig. 1 is fabricated of sheet aluminum.[8] The shallow pyramid design persuades hail to bounce away from the sensor so that multiple hits from a single hailstone are not erroneously recorded. The bottom side contains a piezoelectric ceramic, mounted on a plate with a waterproof cover. A 1-MΩ resistor is placed in parallel with the ceramic disk to bleed off excess charge buildup when not connected to a circuit load. This design was chosen for its low cost, simplicity, and extreme durability in the launch pad environment. However, the greatest challenge in measuring hail is discriminating it from rain.

2. Math Model

The NASA/KSC hail impact transducer is essentially a stiff, sheet aluminum drumhead, where the primary resonance modes are centered on a 4-kHz band. There are numerous modes of vibration, closely spaced in frequency and broad enough in bandwidth to be considered a continuum. However, for the purpose of this analysis, the response of the transducer will be modeled as a set of N uncoupled parallel second-order systems (mass-damper-spring), each characterized by a resonant frequency $\omega_0 = \sqrt{\kappa/M}$, damping factor $d = 2\gamma M$, and modal mass M. The equation of motion of a single mode is therefore

$$\ddot{x} + 2\gamma\dot{x} + \omega_0^2 x = F(t)/M, \tag{1}$$

where $F(t)$ is the force of impact from a hydrometeor, and γ is proportional to the damping factor, introduced to simplify the solution to Eq. (1). Since the complete transducer system consists of N uncoupled second-order mechanical oscillators, as depicted in Fig. 2, important transducer characteristics can be deduced by studying the characteristics of a single mass-

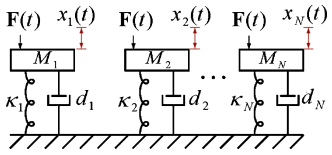


Fig. 2. (Color online) Lumped parameter modal transducer model.

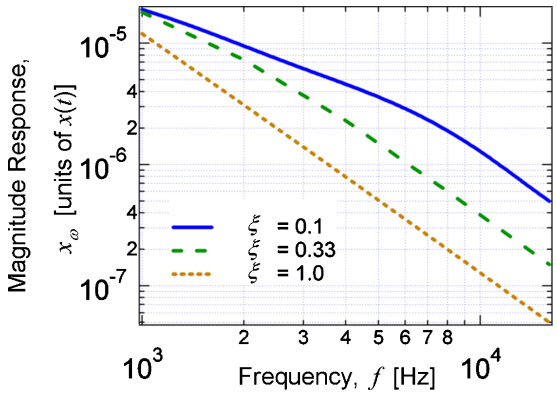


FIG. 3. (Color online) Model response versus frequency for various values of hydrometeor hardness. Movie 1. Frequency response of hail versus rain (video). The plot vertical axis is the magnitude (in dB) of the FFT of the hail transducer signal for single hailstone and rain drop hits. The horizontal axis represents frequency in Hz.

damper-spring system, corresponding to an eigenvalue-eigenvector (or mode) of the vibrating membrane.

A key to solving Eq. (1) for the displacement $x(t)$ is to choose a reasonably simple yet useful description of the impulse force $F(t)$.[9] A reasonable estimate is

$$F(t) = \begin{cases} F_0 = mv(1+\eta)/\tau & t \leqslant \tau \\ 0 & t > \tau \end{cases}, \tag{2}$$

where m is the hydrometeor mass, v is the hydrometeor velocity (usually a constant *terminal velocity*), η is the *coefficient of elasticity*, and τ is the time interval that the impulse force is nonzero. If hail is hard (i.e., not composed of slush), the coefficient of elasticity will be close to the maximum possible value of 1. The coefficient of elasticity can be determined by direct measurement of the velocity v before impact and the reflected velocity v' after impact (assuming one-dimensional movement for simplicity): $v' = -\eta v$, where the minus sign signifies that the final velocity is in the opposite direction.

Taking a very simplistic approach, the time interval τ that the impulse force acts on the transducer can be modeled as

$$\tau = \xi D/v, \tag{3}$$

where D is the hydrometeor diameter, v is the initial velocity before impact, and ξ is an empirical *coefficient of hardness*, which like η, varies between 0 and 1. One might be tempted to relate η to ξ by something like $\xi = 1 - \eta^B$, where B is a fitting parameter. However, the main point here is that η and ξ behave as additive inverses, i.e., when η is near 1, ξ is near 0.

Applying the estimates from Eq. (2) and Eq. (3) to Eq. (1), the solution for displacement is

TABLE I. Parameters value estimates used to generate Fig. 3 with Eq. (4).

γ (s^{-1})	η	M (kg)	D (m)
150	0.9	0.01	0.005

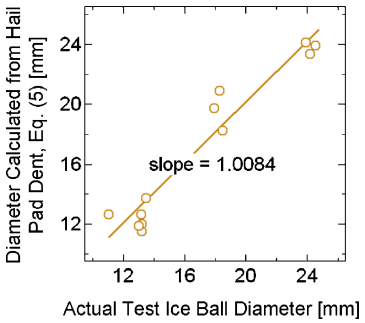


FIG. 4. (Color online) Hail pad calibration check using 20-m drop test data.

$$x(t)=\frac{F_0}{M\omega_0^2}\begin{cases}1-e^{-\gamma t}\left(\cos\omega_1 t+\dfrac{\gamma}{\omega_1}\sin\omega_1 t\right) & t\leqslant\tau\\ e^{-\gamma(t-\tau)}\left(\cos\omega_1(t-\tau)+\dfrac{\gamma}{\omega_1}\sin\omega_1(t-\tau)\right)-e^{-\gamma t}\left(\cos\omega_1 t+\dfrac{\gamma}{\omega_1}\sin\omega_1 t\right) & t>\tau\end{cases}, \tag{4}$$

where $\omega_1\equiv\sqrt{\omega_0^2-\gamma^2}$.

Equation (4) essentially shows that the system impulse response to a finite-width delta function does not decrease with decreasing pulse width as long as the period of the transducer response oscillation is much less than the width of the impulse, $2\pi/\omega_0\ll\tau$. The response roll-off slope above a characteristic frequency becomes independent of hardness, where the offset of the slope is controlled by the impulse time interval τ, from Eq. (2) and Eq. (3), and is inversely proportional to the coefficient of hardness ξ.

The next step is to run some typical values through the second-order system model of Eq. (4), using approximations for hydrometeor terminal velocity[10] and mass, based on hydrometeor diameter D: $v\approx 142D^{1/2}$ and $m=\pi D^3\rho/6$, where ρ is the density of ice. Using the values from Table I in Eq. (4) results in the curves of Fig. 3, corresponding to the transducer signal response, proportional to the displacement $x(t)$, versus mode frequency $f_0=\omega_0/2\pi$, for various values of ξ from $\xi\approx 1$ (raindrop) to $\xi\approx 0.1$ (hard hail). Note that the *magnitude response* x_ω, along the vertical axis in Fig. 3 is defined as the maximum peak value of the magnitude of $x(t)$ in Eq. (4), during the impulse interval, as a function of mode frequency $x_\omega\equiv\max\{|x(t,\omega)|_{t=\text{pulse start}}^{\text{pulse end}}\}$.

The hydrometeor diameter D used in Table I corresponds to a small *rice-size* hailstone, or equivalently, a large raindrop. The damping factor γ is picked to match the decay rate seen in typical time domain plots of the transducer response. When $t=\gamma^{-1}$ or 6.67 ms, the pulse has

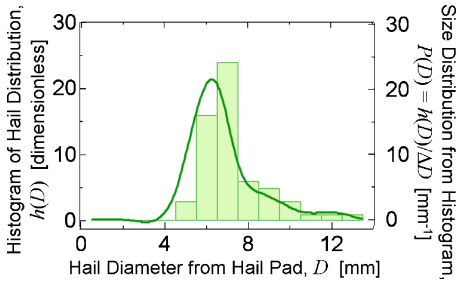


FIG. 5. (Color online) Hail diameters corresponding to hail pad for July 7, 2004, event. Solid line is PDF with $\Delta D=1.0$ mm; bars are histogram.

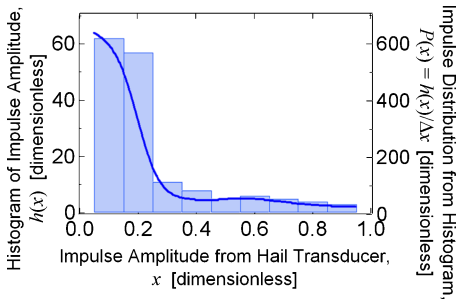


FIG. 6. (Color online) Hail transducer impulse amplitudes corresponding to July 7, 2004, event. Solid line is PDF with $\Delta x = 0.1$; bars are histogram.

decayed to $1/e$ of some initial maximum value. The value chosen for the second-order system mass M has no real meaning here since it is just a scale factor in front of Eq. (4). The system gain of the processing system electronics has a similar effect in the model as the modal mass M. Both effects can be ignored as long as the parameters remain constant.

As discussed previously, the plot in Fig. 3 shows that as mode frequency increases, the separation in response between small and large values of ξ increases. This result is corroborated by the spectral plots of the video in Movie 1, based on the 1024-point fast Fourier transform of a typical hailstone and a raindrop. In the case of hail (i.e., for $\xi < 0.3$), the transducer response in the frequency band above 10 kHz is much higher than for raindrops of similar momentum. This then is the key to rain-versus-hail discrimination when using this type of drop impact transducer. The method that was chosen to process the transducer signal is based on a second-order IIR high-pass filter with a cutoff of 16 kHz. It is important to note that this implies that a sufficient sample rate of at least 44 kHz or better must be used in a digital implementation of the filter.

Mm 1 Frequency response of hail versus rain (video). (.mpg movie file 2.19 MB)

3. Calibration Strategy

A hail pad calibration method developed during the 1978 Alberta Hail Project relates the dent diameter d to the hail diameter D as an empirical second-order polynomial,[11]

$$D = a_0 + a_1 d + a_2 d^2, \tag{5}$$

where $a_0 = 0.38$ cm, $a_1 = 1.11$, and $a_2 = -0.04$ cm^{-1}. Note that Eq. (5) implies that the smallest detectable hail size is $\lim_{d \to 0} D = a_0$, or about 4 mm. Any hailstone smaller than a_0 will go undetected by the hail pad.

In order to verify Eq. (5), ice balls were fabricated, measured, and then dropped from a 20-m platform. The ice ball diameters, hail pad dents (Ref. 7), and corresponding diameters

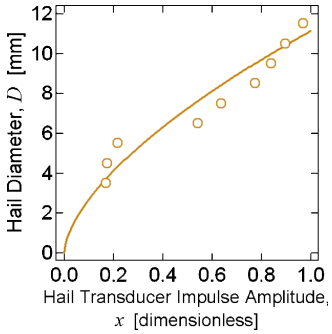


FIG. 7. (Color online) Hail transducer calibration curve based on the July 7, 2004 event. Solid line is fit from Eq. (8) with $A = 11.13$ and $b = 0.621$.

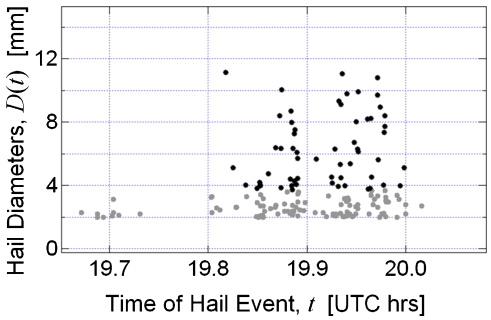


FIG. 8. (Color online) July 7 WEK hail data based on the calibration curve of Eq. (8). Gray circles correspond to extrapolated values below the hail pad D_0 threshold.

calculated from the hail pad calibration formula of Eq. (5) are plotted in Fig. 4 as a function of actual size. Since the slope is very close to unity, Eq. (5) is therefore validated as a means to calibrate the hail transducer using an *in situ*, side-by-side comparison with standard hail pads during naturally occurring hail events. It should be noted that Eq. (5) does not take into account hail hardness. However, experts at the CoCoRaHS analysis lab at CSU are able to estimate hail hardness by observing the type of dent pattern. For the purposes of this current work, hail pad dents, fabricated hail balls, and the resulting calibration all assume hard hail.

Data was collected from a July 7, 2004, thunderstorm containing hail that passed over KSC and the surrounding area. Both the hail transducer and hail pads colocated at three different sites were used to collect data. Site WEK, about 30 km south of KSC, measured the largest amount of hail. The other sites collected only trace amounts. The probability distribution functions (PDFs) of measured hail hits and corresponding discrete histograms for both the hail pads and the hail transducer are shown in Figs. 5 and 6, respectively.

The relationship between the PDF $P(x)$ and histogram $h(x)$ of variable x is $P(x)\Delta x = h(x)$, where Δx is the width of the histogram bin. Each kth bin of $h(x)$ corresponds to x in the range of $(k-1)\Delta x \leqslant x < k\Delta x$. Note that in the case of the hail transducer, x is a dimensionless impulse peak amplitude, similar to x_ω, but not restricted to a narrow-band frequency, scaled to have a value between 0 and 1 [x should not be confused with the mechanical displacement $x(t)$ from previous discussions]. This scaling is convenient for implementation on a fixed-point fractional digital signal processor where numbers are restricted to the range of -1 to 1.

The value of x is dependent on many steps in the signal processing stream and can be related to a physical signal voltage through the conversion of mechanical displacement via the piezoelectric element of the transducer. Because absolute voltage units are not needed when following this type of calibration procedure, it is practical to calibrate the hail monitor system with the assumption that the steps in the signal processing remain constant and have a predictable effect on the impulse amplitude measurement unless an intentional adjustment is made in the signal processing flow.

The PDF of drop diameters from the hail pad can be used to generate a calibration curve for the hail transducer using a probability matching method.[12] Note that this procedure assumes that the number of impulses counted by the hail transducer is greater than the number counted by the hail pad. If the opposite is the case, a slightly modified procedure is followed. Since the hail pad cannot detect hail sizes smaller than $D_0 = a_0 \approx 4$ mm, the initial step in matching PDFs is to solve for the impulse value x_0 corresponding to D_0,

$$\int_{x_n}^{\infty} P(x)dx = \int_{D_n}^{\infty} P(D)dD, \tag{6}$$

with $n=0$, or in the case of the discrete histograms, consisting of M_x and M_D bins,

$$\sum_{k=k_n}^{M_x} h_x(k) = \sum_{j=j_n}^{M_D} h_D(j), \tag{7}$$

where $x_k \equiv k\Delta x$ and $D_j \equiv j\Delta D$. The result of the application of Eq. (7) to the July 7 WEK histograms $h(x_k) \equiv h_x(k)$ and $h(D_j) \equiv h_D(j)$ is $D_0 = 4.61$ mm and $x_0 = 0.169$. This procedure is repeated for all values of j_n and k_n by matching all accumulated histogram totals, bin by bin.

Figure 7 shows the result of matching all histogram bins, corresponding to this *in situ calibration* of the hail transducer. A power-law fit of the form

$$D(x) = Ax^b, \tag{8}$$

where $A = 11.13$ and $b = 0.621$, provides a convenient closed-form calibration formula for all values of impulse amplitude x, consistent with the July 7 WEK data. Figure 8 shows the result of converting the impulse amplitude hail transducer data to equivalent hail diameters using Eq. (8). The open circles correspond to extrapolated values below the hail pad D_0 threshold.

4. Summary

The mathematical model described by Eq. (4) demonstrates that the transducer's magnitude response (maximum of electrical impulse signal) is more sensitive to the hydrometeor hardness at higher frequencies. The result is that discrimination between liquid and solid hydrometeors is best achieved in a frequency band at least two to three times higher than the main transducer resonance. A signal processing implementation to utilize this effect, when sampling at a standard frequency of 48 kHz, simply involves processing the signal with a second-order high-pass filter with a cutoff frequency of 16 kHz. The active impact transducer has the potential to measure hailstone sizes well below the lower size threshold of the passive hail pad.

Future work may include investigating the correlation between impulse frequency spectral response and hail hardness. In addition, incorporating hail hardness into the calibration scheme is an important area of future work. In the case of drop tower calibration, fabricating hail balls with varying degrees of hardness might be accomplished by using small traces of impurities in the water before freezing.[13] Undoubtedly, ignoring hail hardness could lead to large errors in hail size distribution estimates from both the passive hail pad and the active hail transducer.

References and links

[1] J. Joss and A. Waldvogel, "Ein Spektograph fur Niederschlagstopfen mit Automatischer Auswertung," ("A raindrop spectrograph with automatic analysis,") Pure Appl. Geophys. **68**, 240–246 (1967).
[2] J. Lane, T. Kasparis, and G. McFarquhar, "Acoustic Rain Gauge Array Experiment: Phase I," *ERIM 4th International Conference on Remote Sensing for Marine and Coastal Environments*, Orlando, Florida, pp. II-311-II-320 (March 17–19, 1997).
[3] J. Lane, T. Kasparis, and G. McFarquhar, "Adaptive DSP algorithm for calibrating drop size distribution rain gauges," *SPIE AeroSense, 11th International Conference*, Orlando, FL, (June 1997).
[4] B. Federer and A. Waldvogel, "Hail and raindrop size distributions from a Swiss multicell storm," J. Appl. Meteorol. **14**, 91–97 (1975).
[5] E. P. Lozowski and G. S. Strong, "On the calibration of hailpads," J. Appl. Meteorol. **17**, 521–528 (1978).
[6] W. E. Bardsley, "On the maximum observed hailstone size," J. Appl. Meteorol. **29**, 1185–1187 (1990).
[7] Henry Reges, "NASA's hail monitors tested with CoCoRaHS observers," *The Gauge*, The CoCoRaHS Network Newsletter, **3**(1) (Fall 2005), p. 2, http://www.cocorahs.org
[8] Robert Youngquist *et al.*, "A study of hail monitoring strategies for shuttle launch safety," *Research and Technology 2003 Annual Report*, John F. Kennedy Space Center, 8–9 (2003).
[9] R. A. Becker, *Introduction to Theoretical Mechanics* (McGraw-Hill, New York, 1954), pp. 173–179.
[10] R. Gunn and G. D. Kinzer, "The terminal velocity of fall for water droplets in stagnant air," J. Appl. Meteorol. **6**, 243–248 (1948).
[11] G. S. Strong and E. P. Lozowski, "An Alberta study to objectively measure hailfall intensity," Atmos. Oceanic Phys. **4**, 33–53 (1977).
[12] D. Rosenfeld, D. B. Wolf, and D. Atlas, "General probability-matched relations between radar reflectivity and rain rate," J. Appl. Meteorol. **32**, 50–72 (1993).
[13] I. H. Baile and W. C. Macklin, "The effects of impurities on the mechanical strength of accreted ice," J. Atmos. Sci. **24**, 707–710 (1967).

ACOUSTICAL NEWS—USA

Elaine Moran
Acoustical Society of America, Suite 1NO1, 2 Huntington Quadrangle, Melville, NY 11747-4502

***Editor's Note**: Readers of the journal are encouraged to submit news items on awards, appointments, and other activities about themselves or their colleagues. Deadline dates for news items and notices are 2 months prior to publication.*

New Fellows of the Acoustical Society of America

J. Gregory McDaniel
For contributions to structural acoustics

Charles M. Salter
For contributions to the teaching of architectural acoustics and to its practical applications

Ralph Stephen
For contributions to seafloor elastic wave propagation

The 150th meeting of the Acoustical Society of America held in Minneapolis, Minnesota

The 150th meeting of the Acoustical Society of America was held 17–21 October 2005 at the Hilton Minneapolis Hotel in Minneapolis, Minnesota. The 21st annual conference of the Institute of Noise Control Engineering, NOISE-CON 2005, was held concurrently with the ASA meeting from 17 to 19 October. This was the second time that the Society has met in this city, the previous meeting being held in 1984.

The joint meeting drew a total of 1110 registrants, including 122 nonmembers and 238 students. There were 115 registrants (that is, 10%) from outside North America including 18 from Japan, 14 from Korea, 11 each from Germany and the UK, 6 from France, 4 from the Netherlands, 3 from Italy, 2 each from Finland, Norway and South Africa, and 1 each from Australia, Brazil, China, Costa Rica, Denmark, Hong Kong, Israel, Portugal, Sweden, and Taiwan. North American countries, Canada, Mexico, and the United States, accounted for 28, 2, and 995, respectively.

A total of 779 papers, organized into 100 sessions, covered the areas of interest of all 13 Technical Committees of the ASA and NOISE CON 2005. NOISE CON 2005 included 38 technical sessions, 35 of which were cosponsored by the ASA Technical Committee on Noise. There were also several technical sessions cosponsored by NOISE CON with other ASA Technical Committees including Architectural Acoustics, Physical Acoustics, Signal Processing in Acoustics and Structural Acoustics and Vibration.

The meeting also included 13 meetings dealing with standards. Special technical events included a short course, tutorial lecture, distinguished lecture, and technical tours. A short course on "Statistical Energy Analysis" given by Phil Shorter of ESI US R&D was held on 16 and 17 October and attended by 22 participants. The Monday evening tutorial lecture series was continued by Carr Everbach, Swarthmore College. His tutorial "Diagnostic Imaging in Biomedical Ultrasound" was presented to an audience of about 65. Manfred Schroeder, University of Gottingen, presented the Distinguished Lecture titled "From Philharmonic Hall to number theory: The way to more diffusion." The local committee arranged technical tours to Aero Systems Engineering and to a rehearsal of the famed Minnesota Orchestra. This was followed by a discussion with Osmo Vanska, conductor of the orchestra and a guided tour of Orchestra Hall with Cyril Harris as tour guide.

The Society's 13 Technical Committees held open meetings during the Minneapolis meeting where they made plans for special sessions at upcoming ASA meetings, discussed topics of interest to the attendees and held informal socials after the end of the official business. These meetings are working, collegial meetings, and all people attending Society meetings are encouraged to attend and to participate in the discussions. More information about Technical Committees, including minutes of meetings, can be found on the ASA Website ⟨http://asa.aip.org/committees.html⟩ and in the Acoustical News USA section of JASA in the September issue.

An exposition was held in conjunction with the meeting sponsored by ASA and INCE. It included over 45 displays with materials and services for the acoustical and vibration community. It included exhibits of computer-based instrumentation, sound level meters, devices for noise control, sound prediction, among others. The exposition was held from Monday through Wednesday noon and began with an opening reception on Monday evening.

Two special workshops were held at the meeting. The ASA Student Council and the Committee on Women in Acoustics organized a "Mentoring Session for Early Career Acousticians," which included roundtable discussions on topics including early academic careers, balancing family and career, interviewing for jobs, and finding/handling postdoctoral research positions. A special workshop session was provided for ASA journal authors and readers during which a representative from the American Institute of Physics described the electronic services provided to ASA journal authors and readers including an overview of the Peer Xpress manuscript submission and

 0001-4966/2006/119(3)/1281/8/$22.50

FIG. 1. ASA President William Yost (r) presents the 2004 Science Writing Award in Acoustics for Journalists to Kate Ramsayer (l)

handling system. Other services provided through AIP's *Scitation* service were also described.

Social events included the two social hours held on Tuesday and Thursday, a reception for students, the Fellows Luncheon, the Women in Acoustics luncheon, and the morning coffee breaks. These social events provided the settings for participants to meet in relaxed settings to encourage social exchange and informal discussions. A program for accompanying persons included a daily hospitality room where information about activities in and around Minneapolis was provided.

Mildred Dresselhaus, Professor of Physics and Engineering at the Massachusetts Institute of Technology, was the speaker at the Fellows luncheon, which was attended by over 130 people. The Fellows luncheon is now open to all meeting attendees. The Women in Acoustics Luncheon was held on Wednesday afternoon and was attended by over 85 people. A special program for students to meet one-on-one with members of the ASA over lunch, which is held at each meeting, was organized by the Committee on Education in Acoustics. The ASA Student Council hosted a Student Reception with about 125 people in attendance.

The plenary session included the NOISE CON 2005 Closing Ceremony, a business meeting of the Acoustical Society, announcements, acknowledgment of the volunteers who organized the meeting and the presentation of awards and certificates to newly-elected Fellows.

The NOISE CON 2005 Closing Ceremony included acknowledgement of the volunteers who organized the NOISE CON portion of the meeting and awarding the Martin Hirschorn IAC prize and Student Paper Prizes. INCE President Gerald Lauchle acknowledged and expressed thanks to the members of the NOISE CON 2005 organizing committee which included Daniel J. Kato, General Chair; Robert J. Bernhard, General Co-Chair; Patricia Davies and J. Stuart Bolton, Technical Program Co-Chairs; and Richard J. Peppin, Exhibition Manager.

The ASA Science Writing Award in Acoustics for Journalists was presented to Kate Ramsayer for her article "Infrasonic Symphony" published in the 10 January 2004 issue of *Science News* (see Fig. 1). *Declan Butler* also received this award but could not be present at the meeting. He was awarded

FIG. 2. ASA President William Yost (r) presents the 2005 Rossing Prize in Acoustics Education to Katherine S. Harris (l)

the second Science Writing Award in Acoustics for Journalists for his article "Sound and Vision" which was published in the 5 February 2004 issue of *Nature*.

The 2005 Rossing Prize in Acoustics Education was presented to Katherine S. Harris, City University of New York (see Fig. 2). Dr. Harris presented the Acoustics Education Prize Lecture titled "Speech neglect: A strange educational blind spot" earlier in the meeting.

FIG. 3. ASA President William Yost (r) presents the Pioneers of Underwater Acoustics Medal to Henrik Schmidt (l)

FIG. 4. ASA President William Yost (r) presents the Silver Medal in Animal Bioacoustics to James A. Simmons (l)

The Pioneers of Underwater Acoustics Medal was presented to Henrik Schmidt of the Massachusetts Institute of Technology "for pioneering contributions of numerical modeling and at-sea experiments in underwater acoustics" (see Fig. 3). The Silver Medal in Animal Bioacoustics was presented to James A. Simmons of Brown University "for contributions to understanding bat echolocation" (see Fig. 4). The Silver Medal in Speech Communication was presented to Katherine S. Harris of the City University of New York "for research and leadership in speech production" (see Fig. 5). The Trent-Crede Medal was presented to Jerry H. Ginsberg of the Georgia

FIG. 5. ASA President William Yost (l) presents the Silver Medal in Speech Communication to Katherine S. Harris (r)

FIG. 6. ASA President William Yost (r) presents the Trent Crede Medal to Jerry H. Ginsberg (l)

Institute of Technology "for contributions to the theory of vibrations of complex systems" (see Fig. 6). Encomiums for the award recipients can be found in the meeting program [J. Acoust. Soc. Am. **118**, 1982–1997 (2005)].

The election of 15 members to Fellow grade was announced. New fellows are: Wolfgang Ahnert, Kim C. Benjamin, Charles C. Church, Robert L. Clark, James P. Cottingham, C. Craig Formby, Samir N.Y. Gerges, Sheryl Gracewski, Kenneth W. Grant, Paul Johnson, James M. Kates, Qing Huo Liu, Peter Mapp, Peggy B. Nelson, and Jixun Zhou (see Fig. 7). Each new fellow was presented with a certificate and an ASA Fellows pin.

ASA President William Yost expressed the Society's thanks to the Local Committee for the excellent and meticulous planning and operation of the meeting. He introduced Peggy Nelson (see Fig. 8), Chair of the 150th meeting, who acknowledged the contributions of the members of her committee including: Neal F. Viemeister, Technical Program Chair; Kay Hatlestad, Food Service/Social Events; Bruce Olson, Audio-Visual; David Braslau, Accompanying Persons Program/Technical Tours; Derrick Knight and Benjamin Munson, Signs/Publicity. She also expressed thanks to the members of the Technical Program Organizing Committee: Neal F. Viemeister, Tech-

FIG. 7. New Fellows of the ASA

FIG. 8. Peggy Nelson

nical Program Chair; James F. Lynch, T. Martin Siderius, Acoustical Oceanography; Michael J. Ferragamo, Animal Bioacoustics; David Braslau, Bruce C. Olson, Architectural Acoustics; Emad S. Ebbini, Biomedical Ultrasound/Bioresponse to Vibration; Ian M. Lindevald, Education in Acoustics and Musical Acoustics; Stephen C. Thompson, Engineering Acoustics, Michael R. Stinson, Noise; J. Stuart Bolton, Patricia Davies, NOISE CON 2005; James P. Chambers, Physical Acoustics; Magdalena Wojtczak, Psychological and Physiological Acoustics; David M. Fromm, Charles F. Gaumond, Signal Processing in Acoustics; Arlene E. Carney, Benjamin Munson, Speech Communication; Courtney B. Burroughs, Structural Acoustics and Vibration; Dezhang Chu, David R. Dowling, Underwater Acoustics; Julie Golias, TPOM Assistant.

The full technical program can be found in the printed meeting program or online at ⟨scitation.aip.org/JASA⟩ for readers who wish to obtain further information about the Minneapolis meeting. Also, some presenters have provided their full papers or presentation materials for downloading on the ASA Meeting Papers website at ⟨http://scitation.aip.org/asameetingpapers/top.jsp⟩.

We hope that you will consider attending a future meeting of the Society to participate in the many interesting technical events and to meet with colleagues in both technical and social settings. Information about future meetings can be found in the *Journal* and on the ASA Home Page at ⟨http://asa.aip.org⟩.

WILLIAM A. YOST
President 2005–2006

USA Meetings Calendar

Listed below is a summary of meetings related to acoustics to be held in the U.S. in the near future. The month/year notation refers to the issue in which a complete meeting announcement appeared.

2006

6–9 June	151st Meeting of the Acoustical Society of America, Providence Rhode Island [Acoustical Society of America, Suite 1NO1, 2 Huntington Quadrangle, Melville, NY 11747-4502; Tel.: 516-576-2360; Fax: 516-576-2377; Email: asa@aip.org; WWW: http://asa.aip.org].
17–21 Sept.	INTERSPEECH 2006 (ICSLP 2006), Pittsburgh, PA [www.interspeech2006.org ⟨http://www.interspeech2006.org/⟩]
28 Nov–2 Dec	152nd Meeting of the Acoustical Society of America joint with the Acoustical Society of Japan, Honolulu, Hawaii [Acoustical Society of America, Suite 1NO1, 2 Huntington Quadrangle, Melville, NY 11747-4502; Tel.: 516-576-2360; Fax: 516-576-2377; Email: asa@aip.org; WWW: http://asa.aip.org]. Deadline for receipt of abstracts: 30 June 2006.

2007

4–8 June	153rd Meeting of the Acoustical Society of America, Salt Lake City, Utah [Acoustical Society of America, Suite 1NO1, 2 Huntington Quadrangle, Melville, NY 11747-4502; Tel.: 516-576-2360; Fax: 516-576-2377; Email: asa@aip.org; WWW: http://asa.aip.org].
27 Nov–2 Dec	154th Meeting of the Acoustical Society of America, New Orleans, Louisiana (note Tuesday through Saturday) [Acoustical Society of America, Suite 1NO1, 2 Huntington Quadrangle, Melville, NY 11747-4502; Tel.: 516-576-2360; Fax: 516-576-2377; Email: asa@aip.org; WWW: http://asa.aip.org].

2008

28 July–1 Aug	9th International Congress on Noise as a Public Health Problem (Quintennial meeting of ICBEN, the International Commission on Biological Effects of Noise). Foxwoods Resort, Mashantucket, CT [Jerry V. Tobias, ICBEN 9, Post Office Box 1609, Groton, CT 06340-1609, Tel. 860-572-0680; Web: www.icben.org. Email icben2008@att.net.

Cumulative Indexes to the Journal of the Acoustical Society of America

Ordering information: Orders must be paid by check or money order in U.S. funds drawn on a U.S. bank or by Mastercard, Visa, or American Express credit cards. Send orders to Circulation and Fulfillment Division, American Institute of Physics, Suite 1NO1, 2 Huntington Quadrangle, Melville, NY 11747-4502; Tel.: 516-576-2270. Non-U.S. orders add $11 per index. Some indexes are out of print as noted below.

- **Volumes 1–10, 1929–1938:** JASA, and Contemporary Literature, 1937–1939. Classified by subject and indexed by author. Pp. 131. Price: ASA members $5; Nonmembers $10
- **Volumes 11–20, 1939–1948:** JASA, Contemporary Literature and Patents. Classified by subject and indexed by author and inventor. Pp. 395. Out of Print
- **Volumes 21–30, 1949–1958:** JASA, Contemporary Literature and Patents. Classified by subject and indexed by author and inventor. Pp. 952. Price: ASA members $20; Nonmembers $75
- **Volumes 31–35, 1959–1963:** JASA, Contemporary Literature and Patents. Classified by subject and indexed by author and inventor. Pp. 1140. Price: ASA members $20; Nonmembers $90
- **Volumes 36–44, 1964–1968:** JASA and Patents. Classified by subject and indexed by author and inventor. Pp. 485. Out of Print.
- **Volumes 36–44, 1964–1968:** Contemporary Literature. Classified by subject and indexed by author. Pp. 1060. Out of Print
- **Volumes 45–54, 1969–1973:** JASA and Patents. Classified by subject and indexed by author and inventor. Pp. 540. Price: $20 (paperbound); ASA members $25 (clothbound); Nonmembers $60 (clothbound)
- **Volumes 55–64, 1974–1978:** JASA and Patents. Classified by subject and indexed by author and inventor. Pp. 816. Price: $20 (paperbound); ASA members $25 (clothbound); Nonmembers $60 (clothbound)
- **Volumes 65–74, 1979–1983:** JASA and Patents. Classified by subject and indexed by author and inventor. Pp. 624. Price: ASA members $25 (paperbound); Nonmembers $75 (clothbound)
- **Volumes 75–84, 1984–1988:** JASA and Patents. Classified by subject and indexed by author and inventor. Pp. 625. Price: ASA members $30 (paperbound); Nonmembers $80 (clothbound)
- **Volumes 85–94, 1989–1993:** JASA and Patents. Classified by subject and

indexed by author and inventor. Pp. 736. Price: ASA members $30 (paperbound); Nonmembers $80 (clothbound)

- **Volumes 95–104, 1994–1998:** JASA and Patents. Classified by subject and indexed by author and inventor. Pp. 632. Price: ASA members $40 (paperbound); Nonmembers $90 (clothbound)
- **Volumes 105–114, 1999–2003:** JASA and Patents. Classified by subject and indexed by author and inventor. Pp. 616. Price: ASA members $50; Nonmembers $90 (paperbound)

Revision List

New Associates

Anderson, Van B., Monitor Instruments, Inc., 437 Dimmocks Mill Rd., Hillsborough, NC 27278

Azar, Lawrence, PPB. Inc., 740 13th St., Ste. 325, San Diego, CA 92101

Bang, John Y., 1910 Glenview Ave., Park Ridge, IL 60068

Bassar, Christopher J., 264 N. Cypress Ave., Fresno, CA 93727

Bibb, John S., Bldg. Construction Products, Caterpillar, Inc., 100 Regency Forest Dr., Ste. 300, Cary, NC 27511

Boldt, Jeffrey G., KJWW Engineering Consultants, 802 W. Broadway, Madison, WI 53713

Buzduga, Mariana, Scantek, Inc., Calibration Lab., 7060 Oakland Mills Rd., Ste. L, Columbia, MD 21046

Cao, Zhengliang, Hangzhou Applied Acoustics Research Inst., State Key Lab. of Oceanic Acoustics, Rm. 1510, No. 96, HuaXing Rd., Hangzhou, Zhejiang 310012, China

Clemins, Patrick J., 2557 N. Terrace Ave., Apt. A, Milwaukee, WI 53211

de Boer, Egbert, Oldambtstraat 7, Amsterdam, NH 1079 PT, The Netherlands

Djurek, Ivan, Vladimira Ruzdjaka 17, Zagreb HR 10000, Croatia

Ekimov, Alexander, Univ. of Mississippi, National Ctr. for Physical Acoustics, 1 Coliseum Dr., University, MS 38677

Farassat, Fereidoun, Aeroacoustics Branch, NASA Langley Research Ctr., MS 461, 2 N. Dryden St., Hampton, VA 23681-2291

Farbo, Aaron M., Stewart Acoustical Consultants, 7406L Chapel Hill Rd., Raleigh, NC 27607

Ferguson, Glen, Benthos, Inc., 49 Edgerton Dr., North Falmouth, MA 02556

Fielding, Joseph T., Morgan Electro Ceramics, 232 Forbes Rd., Bedford, OH 44146

Fraiman, Gary, Scantek Inc., Sales & Service, 7060 #L Oakland Mills Rd., Columbia, MD 21046

Furihata, Kenji, Dept. of Electrical and Electronic Eng., Shinshu Univ., Wakasato 4-17-1, Nagano 380-8553, Japan

Giddens, Eric M., Scripps Inst. of Oceanography, Univ. of California, San Diego, 8820 Shellback Way, La Jolla, CA 92037-0238

Hardiman, James E., Lockheed Martin, Underwater Vehicles, 7 Barnabas Rd., Marion, MA 02738

Harney, Kieran P., Micromachined Products Division, 21 Osborn St., Cambridge, MA 02139

Hermann, Heinrich G., Hermann Design Studio, 21 Arrowhead Rd., Concord, MA 01742

Holtrop, James S., Acousticontrol LLC, 1709 Washington Ave., Loft 300, St. Louis, MO 63103

Huang, Lixi, Mechanical Engineering, The Hong Kong Polytechnic Univ., Kowloon, China

Hughes, Charles E., Excelsior Audio Design and Services, LLC, 2130 Shadwell Ct., Gastonia, NC 28056

Hurley, David H., Idaho National Lab., P.O. Box 1625, Idaho Falls, ID 83415-2209

Karas, Eric L., Environment and Regeneration, London Borough of Islington, 159 Upper St., London N1 1RE, UK

Kassewitz, Jack, SpeakDolphin.com, 7980 SW 157th St., Miami, FL 33157

Kayan, Thomas E., 330 Wade St., Bridgeport, CT 06604

Kim, Hee-Sun, Dept. of Asian Languages, Stanford Univ., Main Quad Bldg. 50, Stanford, CA 94305

Kim, Benjamin J., 964 Towlston Rd., McLean, VA 22102

Knox, Ryan, Acoustic Dimensions, 15505 Wright Brothers Dr., Addison, TX 75001

Kob, Malte, RWTH Aachen Univ., Dept. of Phoniatrics, Pedaudio. & Commun. Disorders, Pauwelsstr. 30, Aachen, NRW 52074, Germany

Kritski, Alexander, Statoil Research Centre, Arkitekt Ebbellsvei 10, Trondheim 7005, Norway

Kuthirummal, Narayanan, Physics and Astronomy, College of Charleston, 58 Coming St., Charleston, SC 29424

Langthjem, Mikael A., Dept. of Mechanical Systems Engineering, Yamagata Univ., Faculty of Engineering, Jonan 4-3-16 Yonezawa, Yamagata 992-8510, Japan

Lee, Iljae, Ctr. for Automotive Research, Ohio State Univ., 930 Kinnear Rd., Columbus, OH 43212

Leung, Wai Kit, Bose Corporation, Automotive System Division, 688 Great Rd., P.O. Box 9102, Stow, MA 01775-9102

Lewis, Russ, 2404 9 Ave., NW, Calgary, AB T2N 1E7, Canada

Lin, Weili, Starkey Labs Inc., 6600 Washington Ave., S., Eden Prairie, MN 55344

Lipscy, Virginia A., Springer, 233 Spring St., New York, NY 10013

Liu, Sheng, Sound ID, 3430 W. Bayshore Rd., Palo Alto, CA 94303

Mahn, Jeffrey, Black & Decker, 701 E. Joppa Rd., TW080, Towson, MD 21030

McArdle, Rachel A., 13189 87th Ave., N., Seminole, FL 33776

McKee, Matthew M., GN ReSound, 2601 Patriot Blvd., Glenview, IL 60026

McPherson, Michael S., 1903 Anderson Rd., A-2, Oxford, MS 38655

Murray, James J., Lockheed Martin Orincon, 4350 N. Fairfax, Arlington, VA 22203

Ng, Kok-Hwee, Stryker Development, One Broadway, 16th Floor, Cambridge, MA 02142

O'Sickey, Brian, Artisan Sound Control, LLC, 1164 Lloyd Rd., Wickliffe, OH 44092

Perrey, Christian F., GE Medical Systems Kretzlechnik, R&D, Tiefenbach 15, Zipf, OOE 4871, Austria

Perron, Scott M., The Soundcoat Company, Inc., One Burt Dr., Deer Park, NY 11729

Philipp, Norman H., Yantis Acoustical Design, 720 Olive Way, Ste. 1400, Seattle, WA 98101-1853

Pitt, Stephen P., 11410 Whitewing Ave., Austin, TX 78753

Qian, Jinyu, 204 White Pine Cir., Lawrenceville, NJ 08648

Rodriguez, Ramirez, Zeck Audio Iberica S.L Poligono Industrial Barrafuerte naves 12-13, C.P. 30880. Apt de correos 219, Aguilas, (Murcia), Spain

Rogers, Robert E., Audio/Video Designs, 216 N. Broadway, Moore, OK 73169

Roper, Joshua M., Charles M. Salter Associates, 325 S. First St., Ste. 160, San Jose, CA 95113

Salzberg, David H., SAIC, Ocean Sciences Division (M/S 1-11-15), 1710 Saic Dr., Mclean, VA 22102

Sauro, Ron, NWAA Labs, Inc., 25132 Rye Canyon Loop, Santa Clarita, CA 91355

Saussu, Patrick T., GN ReSound, 2601 Patriot Blvd., Glenview, IL 60026

Schmitz, Georg, Ruhr-Universitat Bochum, Medical Engineering, Bldg. IC 6, Universitatsstr. 150, Bochum NRW 44780, Germany

Shilton, Simon J., Acustica Limited, Trident One, Styal Rd., Manchester, Cheshire, M22 5XB, UK

Smith, Caroline, Mathematics and Statistics, James Madison Univ., MSC 7803, Burruss Hall, Harrisonburg, VA 22807

Song, Aijun, Univ. of Delaware, College of Marine Studies, Robinson Hall, Newark, DE 19716

Souslov, Alexei V., National High Magnetic Field Lab., 1800 E. Paul Dirac Dr., Tallahassee, FL 32310

Stocklin, Norma L., 3203 Fleet Landing Blvd., Atlantic Beach, FL 32233

Strong, Thomas D., 7580 Varsit Ln., Bremerton, WA 98311

Talayman, Turker S., TeamFORES Acoustics & Electronics Ltd., Cincarli Sk. Karabekir A. No, 446/15 Suadiye, Istanbul 334744, Turkey

Tech, Olaf, Nelkenstr. 7, Recke D-49509, Germany

Teixeira, Erik C., 8814 Los Arboles Ave., NE, Albuquerque, NM 87112

Throckmorton, Chandra, 3209 Hursey St., Durham, NC 27703

Wahlberg, Magnus K. G. Naturens Stemme, Ewaldsgade 4, 2. sal, Aarhus 800, Denmark

Wang, Yingmin, Northwestern Polytechnical Univ., College of Marine Engineering, 127 You Yi Xi Lu, Xi'an Shaanxi, 710072, China

Washburn, Donald J., The Audio Bug, Inc., 1180 NE 165th St., Miami Beach, FL 33162-3847

Watson, Peter J., Speech Language Hearing Sciences, Univ. of Minnesota, 164 Pillsbury Dr., SE, Shevlin Hall 115, Minneapolis, MN 55455

Worley, Wade R., 2218 Greyburn Ln., Houston, TX 77080
Zinn, Ben T., Georgia Inst. of Technology, Aerospace Engineering, 270 Ferst Dr., Atlanta, GA 30332-0150

New Students

Adam, Brittany A., Union College, 807 Union St., B0002, Schenectady, NY 12308
Agus, Trevor R., 3/2, 9 Hanson Park, Glasgow G31 2HJ, Scotland
Alfaro, Josue A., Physical Sciences, Univ. of California, Irvine, Irvine, CA 92697
Baldwin, Steven L., Washington Univ. in St. Louis, Physics, One Brookings Dr., St. Louis, MO 63130
Bay, Mert, 1002 W. Clark St., Apt. 201, Urbana, IL 61801
Bohn, Sandra E., Univ. of Southern Mississippi, Psychology, 118 College Dr., Box 5025, Hattiesburg, MS 39406
Bradley, Rose E., 599 W. Westfield Blvd., Indianapolis, IN 46208
Brannstrom, Jonas, Amiralsgatan 56B, Malmo, Skane 21437, Sweden
Brode, Andrew M., 899 S. Plymouth Ct., 1002, Chicago, IL 60605
Brown, Sarah, 3403 Medina Ln., Bowie, MD 20715
Campbell, Ian C., 1500 St. Olaf Ave., Northfield, MN 55057
Chang, Natasha A., 1697 Broadway St., #303, Ann Arbor, MI 48105
Chen, Lan, 350 Memorial Dr., #101A, Cambridge, MA 02139
Chen, Li-Jen, 212-11, Halsey Dr., West Lafayette, IN 47906
Choi, Susie, Univ. of Southern California, Dept. of Linguistics, Grace Ford Salvatori 301, Los Angeles, CA 90089
Colvard, Matthew T., 1280 Piedra Morada Dr., Palisades, CA 90272
da Silva, Andrey R., 4635 Clanranald Ave., Apt. 301, Montreal QC H3X 2R8, Canada
Davidson, Sean A., Biomedical and Chemical Eng., Syracuse Univ., 621 Skytop Rd., Syracuse, NY 13244
Davies, Claire, 671 Munich Cir., Waterloo, ON N2V 2L6, Canada
DeWitt, Megan S., 1013 Old Boalsburg Rd., #6, State College, PA 16801
Dieckman, Eric A., Missouri Hall Rm. 408, 809 S. Mulanix, Kirksville, MO 63501
Dixon, Ross, 2400 White Hall Rd., P.O. Box 96, White Hall, MD 21161
Drechsel, James, 4666 S. Meadow View Cir., Murray, UT 84107
Eckert, Dan J., 733 Missouri St., Lawrence, KS 66044
Erickson, Joe, 40 W. 152 Wildwood Dr., Aurora, IL 60506
Farnetani, Andrea, Univ. of Ferrara, Engineering Dept., Via Saragat, 1, Ferrara, 44100, Italy
Fuller, Melissa Y., P.O. Box 235417, Honolulu, HI 96823
Gai, Yan, Biomedical and Chemical Engineering, Syracuse Univ., Inst. for Sensory Research, 621 Skytop Rd., Syracuse, NY 13244-5290
German, James B., Northwestern Univ., Linguistics, 2016 Sheridan Rd., Evanston, IL 60208-4090
Gibbs, Bobby E., 25 Brunswick Rd., 1st Floor, Troy, NY 12180
Gregg, Justin D., School of Psychology, Trinity College Dublin, College Green, Dublin 2, Ireland
Hammond, Terri D., National Security Technology, Johns Hopkins Univ., 11100 Johns Hopkins Rd., Laurel, MD 20723-6099
Havens, Nicholas P., Audio Arts and Acoustics, Columbia College Chicago, 600 S. Michigan, Chicago, IL 60606
Hay, Todd A., 1213 Kinney Ave., Austin, TX 78704
Herbison, Sarah W., 251 10th St., NW, #B704, Atlanta, GA 30318
Hiatt, Jessica L., Theer and Associates, 8978 'J' St., Omaha, NE 68127
Hilton, Marshall S., 1144 Louisiana St., #12, Lawrence, KS 66044
Horta, Miguel A., Pennsylvania State Univ., Graduate Program in Acoustics, P.O. Box 30, State College, PA 16804
Hough, Barbara, Colby College, 5860 Mayflower Hill Dr., Waterville, ME 04901
Hwang, Hyekyung, Linguistics, Univ. of Hawaii at Manoa, 569 Moore Hall, 1890 East-West Rd., Honolulu, HI 96822
Ilana, Cellum P., The Graduate Ctr. CUNY, Audiology, 365 Fifth Ave., New York, NY 10016-4309
Ito, Kikuyo, Speech and Hearing Sciences, Graduate Ctr., CUNY, 365 Fifth Ave., New York, NY 10016
Jarick, Michelle A., 111 Larksmere Crt., Markham, ON L3R 3R4, Canada
Jemmott, Colin W., 1730 Bristol Ave., Apt. 622, State College, PA 16801
Jennings, Skyler, Purdue Univ., Speech, Language, Hearing Sciences, Heavilon Hall 500 Oval Dr., West Lafayette, IN 47907-2038
Kamery, Chris, SUNY Fredonia, Physics, 280 Central Ave., Fredonia, NY 14063
Kazemzadeh, Ebrahim F., Computer Science, Univ. Southern California, University Park Campus, Los Angeles, CA 90089
Kesselman, Rachel F., 701 SW 62nd Blvd., Apt. 244, Gainesville, FL 32607
Kim, Byoung-Nam, SungKyunKwan Univ., Dept. of Physics, 300 Chunchun-dong, Suwon, Jangan-gu Gyeonggi-do, 440-746, Republic of Korea
Konefall, Joseph A., 1936 N. Clark St., Chicago, IL 60614
Kornatz, Matthew R., Columbia College, 600 S. Michigan Ave., Chicago, IL 60605
Kuesal, Cenny, 1301 W. 24th St., K-21, Lawrence, KS 66046
Kujawa, Pat C., 329 N. 16th St., Bozeman, MT 59715
Laplanche, Christophe, 20 rue Marcelin Berthelot, Alfortville, 94140, France
Larbi-Cherif, Adrian M., Applied Research Labs., 10000 Burnet, Austin, TX 78758
Lee, Dong-Chang, P.O. Box 750951, Fairbanks, AK 99775-0951
Lee, Jimin, 4723 Sheboygan Ave., #212, Madison, WI 53705
Lee, Jaehyung, Penn State Univ., Acoustics, Graduate Program in Acoustics, P.O. Box 30, State College, PA 16804
Lefkowitz, Kimberly A., 1051 Teaberry Ln., State College, PA 16803
Li, Bin, 332 Univ. Village, S., #1, Gainesville, FL 32603
Liu, Dalie, 2212 'R' St., Apt. 16, Lincoln, NE 68503
Lopes, Leonard V., Aerospace Engineering, Pennsylvania State Univ., College Ave., University Park, PA 16802
Lovitt, Andrew W., 601 E. White St., A24, Champaign, IL 61820
Luedke, Dan E., 3226 Bloomfield Ave., West Hartford, CT 06117
Martin, Coleman J., 14 W. Elm, Apt. 807, Chicago, IL 60610
McGarry, Louise P., Cornell Univ., Ocean Resources & Ecosystems Program, 2140 Snee Hall, Ithaca, NY 14853
Meng, Tianzhu, 180 Clinton St., New Bedford, MA 02740
Messersmith, Jessica J., 2701 N. Cotner Blvd., Lincoln, NE 68507
Miller, Andrew N., 1817 E. 29th St., Lawrence, KS 66046
Nagy, Zvonimir, 1058 W. Thorndale Ave., Apt. 2T, Chicago, IL 60660
Napoletano, Brian M., Purdue Univ., Forestry & Natural Resources, 195 Marsteller St., West Lafayette, IN 47906
Oglesbee, Eric, 5390 W. Hillcrest Dr., Bloomington, IN 47404
Ortallono, Sam, 1258 Commonwealth Ave., #11, Allston, MA 02134
Parakkal, Santosh, Physics and Astronomy, Univ. of Mississippi, 108 Lewis Hall, University, MS 38677
Paumen, Allison L., 6658 75th St., NW, Maple Lake, MN 55358
Pedersen, Carol A., P.O. Box 5946, West End QLD 4101, Australia
Picinali, Lorenzo, via Belfiore 10, Milano, Lombardia, 20145, Italy
Pinto, Alberto, Via San Gimignano, 4, I-20146 Milan, Italy
Pleckham, Jay M., 4330 W. 206th St., Matteson, IL 60443
Rattandait, Kitti, 421 N. 44th St., Apt. 1318, Lincoln, NE 68503
Robinson, Daniel H., Purdue Univ., Herrick Laboratories, Mechanical Engineering, 140 S. Intramural Dr., West Lafayette, IN 47907
Roggles, Dorea R., Gustavus Adolphus College, 800 College Ave., Saint Peter, MN 56082
Rosenblatt, Lindsay, Millsaps College, 1701 N. State St., Jackson, MS 39210
Roverud, Elin M., 417 2nd Ave., SE Apt. 2, Minneapolis, MN 55414
Rubin, Gordon B., Architectural Acoustics, Rensselaer Polytechnic Inst., 110 8th St., Troy, NY 12180
Satar-Boroujeni, Hamid, 349 Pleasant St., Apt. B1-15, Malden, MA 02148-8109
Saunders, Edwin, 155 Summit St., Apt. 404, Newark, NJ 07102
Schaefer, Justin M., Audio Arts and Acoustics, Columbia College Chicago, 624 S. Michigan Ave., Chicago, IL 60605
Schiller, Noah H., 2252 New Kent Ct., Newport News, VA 23681-2199
Schneider, Lauren G., Univ. of Southern Mississippi, Experimental Psychology, 118 College Dr., Box 5025, Hattiesburg, MS 39406
Schultz, Troy D., 1051 Teaberry Ln., Apt. 0208, State College, PA 16803
Setzler, Gabriel L., 6065 Plainview, Detroit, MI 48228
Sheaffer, Jonathan, House 34/4, Kibbutz Afikim 15148, Israel
Shepherd, Micah R., Brigham Young Univ., C110 ESC BYU, Provo, UT 84602
Shin, Yoon Shik, 1966 Cal Dr., West Lafayette, IN 47907

Shinya, Takahito, Linguistics, Univ. of Massachusetts, Amherst, 226 S. College, Amherst, MA 01003
Siemon, Jordan A., 107 Delaware Ave., Danville, IL 61832
Smith, Noelle J., 17 Gilbert Tr., Atlanta, GA 30308
Spencer, Mychal P., Ray W. Herrick Laboratories, Mechanical Engineering, 140 S. Intramural Dr., West Lafayette, IN 47907-2031
Spicer, Benjamin, 315 Radisson Rd., Pittsburgh, PA 15227
Stargell, Gregory, 2980 Jonesboro Rd., #314, Atlanta, GA 30354
Sprague, Kaitlyn A., 8613 Bluedale St., Alexandria, VA 22308
Stilp, Christian E., Psychology, Univ. of Wisconsin, Madison, 1202 W. Johnson St., Madison, WI 53706
Stulgin, Steven J., 817 S. School St., Lombard, IL 60148
Sun, Yang, 2950 Portage Bay, W., #302, Davis, CA 95616
Tashmukhambetov, Arslan M., 802 Gordon St., Lafayette, LA 70894
Ule, Helen J., 1258 Aubin Rd., Windsor ON N8Y 4E5, Canada
Vander Weide, John M., 9968 Pine Island Dr., Sparta, MI 49345
Velankar, Manish R., 301 SW Lincoln St., Apt. 1203, Portland, OR 97201
Webster, Kevin T., 1697 E. Roberts, #104, Fresno, CA 93710
Whelan, Andrew, 2712 Southwind Cove, Oxford, MS 38655
Whittington, Jarrod E., 463 Fulton St., 3rd Floor, Troy, NY 12180
Williams, Chris M., Acoustics, Columbia College, 600 S. Michigan Ave., Chicago, IL 60605
Winn, Matthew B., Hearing and Speech Sciences, Univ. of Maryland, 0100 Lefrak Hall, College Park, MD 20742
Wohler, Adriene A., 6734 High Rd., Darien, IL 60561
Wonmo, Kang, Engineering Mechanics, Univ. of Nebraska-Lincoln, W317.4 Nebraska Hall, Lincoln, NE 68588-0526
Yamamoto, Yoh, 12124 Knight Krossing Cr., Apt. 103-D, Orlando, FL 32817
Yang, Min, Washington Univ. in St. Louis, Physics, One Brookings Dr., St. Louis, MO 63130
Yang, Jie, Georgia Inst. of Technology, Mechanical Engineering, 801 Ferst Dr.., N.W., Atlanta, GA 30332-0405
Zabovnik, Brian A., 13805 Glen Oaks Place, Edmond, OK 73013
Zaharakis, Alex, 75-72 197th St., Flushing, NY 11366
Zakrzewski, Thomas, 8120 Kristo Ln., Orland Park, IL 60462
Zanartu, Matias, 148 Arnold Dr., West Lafayette, IN 47906
Zeman, John H., Univ. of South Dakota, Human Factors, 414 Clark St., Vermillion, SD 57069
Zhao, Shukui, Biomedical Engineering, Univ. of California, Davis, 451 E. Health Sciences Dr., Davis, CA 95616
Zhou, Xinhui, 4329 Rowalt Dr., Apt. 101, College Park, MD 20740

New Electronic Associates

Asdrubali, Francesco, Universita degli Studi di Perugia, Ingegneria Industriale, Via G. Duranti 67, Perugia, 06125, Italy
Bohlen, Thomas, Inst. of Geosciences, Kiehl Univ., Otto-Hahn-Platz 1, Kiel 24118, Germany
Cholakis, Ernest H., Numerical Sound, P.O. Box 1275, Station K, Toronto ON M4P 3E5, Canada
Cohen-Bacrie, Claude R., SuperSonic Imagine, 510 Rue Rene Descartes, Aix-en-Provence, Bouches du Rhone 13857, France
Collins, David W., 127 S. Harper Ave., Los Angeles, CA 90048
Ding, Shuxue, Dept. of Computer Science, Univ. of Aizu, Tsuruga, Ikki-Machi, Aizu-Wakamatsu, Fukusihoma, 965-8580, Japan
Evitts, Paul M., West Chester Univ., CDIS, 201 Carter Dr., Ste. 417, West Chester, PA 19383
Farrell, Glen E., 1789 Avenida Segovia, Oceanside, CA 92056
Gavelle, Noel, 1 Rue du Sain Sauveur, Foix 09000, France
Gibson, Thomas, 10 Masonbrook Ln., Littleton, MA 01460
Gottinger, Bernd, 39 Forest Place, Fredonia, NY 14063
Heddle, James, James Heddle Pty. Ltd., 2315/180 Grey St., Brisbane QLD 4101, Australia
Hotchkiss, Sean S., 504 Liberty St., Petoskey, MI 49770
Ishikawa, Kiyoshi, Hosei Univ., English Dept., 2-17-1 Fujimi, Chiyoda-ku, Tokyo 102-8160, Japan
Jones, Billy D., Alaska Native Technologies, LLC, 26268 Twelve Trees Ln., Ste. B, Poulsbo, WA 98370
Klaeboe, Ronny, Safety and Environment, Inst. of Transport Economics, Grensesvingen 7, Oslo N-0602, Norway
Lee, Scott G., Intel Corp., 2800 Ctr. Dr., M/S DP2-303, DuPont, WA 98327
Lopez-Vidal, Ramon, Manuel De Falla 26, Rota, Cadiz 11520, Spain
Magnasco, Marcelo O., Ctr. for Studies in Physics and Biology, Rockefeller Univ., 1230 York Ave., New York, NY 10021
Matteson, Samuel E., Dept. of Physics, Univ. of North Texas, P.O. Box 311427, Denton, TX 76203
Mitri, Farid G., Mayo Clinic College of Medicine, Physiology & Biomedical Engineering, 200 First St., SW, Rochester, MN 55905
Moore, David F., EBR Systems Inc., 686 W. Maude Ave., Sunnyvale, CA 94085
Oberfeld, Daniel, Psychologisches Inst., Univ. Mainz, Staudingerweg 9, Mainz 55128, Germany
Pörschmann, Christoph, Fachhochschule Köln, Inst. f. Nachrichtentechnik, Betzdorer Str. 2, Köln 50679, Germany
Qiu, Robert, Tennessee Tech Univ., 115 W. 10th St., Cookeville, TN 38501
Seeber, Bernhard U., Dept. of Psychology, Univ. of California at Berkeley, 3210 Tolman Hall #1650, Berkeley, CA 94720-1650
Snyder, Joel S., VA Boston Healthcare System, Harvard Medical School, Dept. of Psychiatry-116A, 940 Belmont St., Brockton, MA 02301
Sudasna, Jintadhee, Maretron LLC, 9034 N. 234rd Ave., Ste. 13, Phoenix, AZ 85021
Sugahara, Mariko, Doshisha Univ., Dept. of English, Karasuma-Imadegawa, Kamigyo-ku, Kyoto, 602-8580, Japan
White, Richard M., 34 Lonsdale Rd., Bournemouth, Dorset BH3 7LY, UK
Wilson, Willsingh, Wilson-Acousticx, Rinkartstr. 1, Berlin 12437, Germany
Wolfe, Joe, Univ. of New South Wales, Physics, Sydney, NSW 2052, Australia

New Corresponding Electronic Associates

Boonyanant, Phakphoom, 32/84 Parichart Village, Rangsit-Nakon Nayok Rd., Klong 4, Klong Luang, Pathumthani, 12120, Thailand
Ha, Kanglyeol, Physics, 599-1, Daeyon 3-Dong, Nam-Gu, Busan, 608-737, Korea
Moscoso, Ana P., Univ. de Las Americas, Av. Colon 338 Y 6 De Diciembre, Quito, Pichincha, 5932, Ecuador
Phani Kiran, Vadavalli V., Plot No. 128, Phase-2, Vijayawada, Andhra Pradesh 521137, India
Wang, Ding, Inst. of Hydrobiology, Chinese Academy of Science, Dolphin Research Group, 7 Donghu S. Rd., Wuhan, Hubei 430072, China
Yang, Gangyi, Chinese Academy of Science, Apt. 425, Bldg. 24, Fuxingmen wai St., Beijing 100038, China

Reinstated

B. Staple—*Member*
T. R. Beiler, E. P. Honda, C. Q. Howard, H. Paek, E. K. Scheer, J. J. Sidtis—*Associates*
J. F. Abbott, A. Mancuso—*Electronic Associates*

Members Elected Fellows

A. G. Askenfelt, A. U. Case, T. Dau, C. Y. Espy Wilson, J. G. McDaniel, L. A. Miller, B. Mohl, S. S. Narayanan, S. D. Richards, R. A. Stephen, D. Tang, G. R. ter Haar

Associates Elected Members

K. J. Bastyr, K. J. Benoit-Bird, C. Buma, A. L. Carey, D. Chhetri, J. L. Cipolla, D. B. Copeland, M. J. Daley, L. Davidson, S. H. Ferguson, B. W. Gick, J. Guan, R. Hagiwara, J. M. Harte, L. L. Holt, Y. Jiang, S. Keith, R. E. Johnson, J. A. Keefe, G. D. Larson, P. H. Mathuria, J. A. McAteer, A. W. Meyer, P. J. Murphy, T. B. Neilsen, B. K. Newhall, A. A. Parthasarathi, M. Pouplier, C. J. Richie, C. B. Roads, T. L. Rogers, D. Rowan, L. S. Smith, J. W. Solway, G. C. Stecker, R. M. Stern, O. G. Symko, S. L. Thomson, K. D. Van Wyk, R. A. Wright, Y. Xu

Associates to Electronic Associates

M. Bertsch, D. A. Cabrera, A. C. Carballeira, R. Drullman, J. E. Joseph, R. H. Keith, J. Kim, S. A. Kleinknecht, B. R. LaCour, B. W. Lewis, L. Menard, R. A. Moscoso, T. F. Noonan, C. J. Pollock, M. K. Rami, R. M. Su, J. M. Suarez, R. Wakeland, U. T. Zwicker

Electronic Associates to Associates

G. M. Hughes, X. Jiang, E. Rambod, E. L. Zechmann

Students to Associates

R. Aravamudhan, T. C. Bent, N. F. Declercq, G. L. Gattemy, J. Mamou, E. M. Oleson, C. Tollefsen

Students to Electronic Associates

P. B. Allen, G. A. Coudriet, A. Jesse, J. L. Loebach, M. S. Palahanska-Mavrov, A. Salomon, M. Xu, B. Zeitler

Associate to Student

M. A. Andrew

Resigned

R. J. Bobber, G. C. Izenour—*Fellows*
T. H. Fay—*Member*
I. Akio, S. Fischer, O. S. Higgins, C. Rice—*Associates*

Deceased

F. A. Everest, F. J. Fry, J. Hind, J. W. Kopec—*Fellows*
P. Ebaugh, V. P. Garwood, C. G. Gordon, R. V. Lewis, J. R. Mundie, A. Nelkin, H. Schecter—*Members*

Fellows	890
Members	2329
Associates	2774
Students	1071
Electronic Associates	483
	—
	7547

ACOUSTICAL STANDARDS NEWS

Susan B. Blaeser, Standards Manager
ASA Standards Secretariat, Acoustical Society of America, 35 Pinelawn Rd., Suite 114E, Melville, NY 11747 [Tel.: (631) 390-0215; Fax: (631) 390-0217; e-mail: asastds@aip.org]

George S. K. Wong
Acoustical Standards, Institute for National Measurement Standards, National Research Council, Ottawa, Ontario K1A 0R6, Canada [Tel.: (613) 993-6159; Fax: (613) 990-8765; e-mail: george.wong@nrc.ca]

American National Standards (ANSI Standards) developed by Accredited Standards Committees S1, S2, S3, and S12 in the areas of acoustics, mechanical vibration and shock, bioacoustics, and noise, respectively, are published by the Acoustical Society of America (ASA). In addition to these standards, ASA publishes Catalogs of Acoustical Standards, both National and International. To receive copies of the latest Standards Catalogs, please, contact Susan B. Blaeser.

Comments are welcomed on all material in Acoustical Standards News.

This Acoustical Standards News section in JASA, as well as the National and International Catalogs of Acoustical Standards, and other information on the Standards Program of the Acoustical Society of America, are available via the ASA home page: http://asa.aip.org.

Standards Meetings Calendar National

The 151st ASA Meeting, Providence, RI, 5–9 June 2006, the ASA Committee on Standards (ASACOS), ASACOS Steering Committees, and Accredited Standards Committees S1 Acoustics, S2 Mechanical Vibration and Shock, S3 Bioacoustics, and S12 Noise, and the Standards Plenary Group will meet as follows:

• ***Monday, 5 June 2006***

ASACOS Steering Committee.

• ***Tuesday, 6 June 2006***

ASA Committee on Standards (ASACOS). Meeting of the Committee that directs the Standards Program of the Acoustical Society.

• ***Wednesday, 6 June 2006***

Standards Plenary Group. To discuss national and international items relevant to Accredited Standards Committees S1, S2, S3, and S12. This meeting also includes the annual meetings of the U.S. Technical Advisory Groups (TAGs) for ISO/TC 43 Acoustics, ISO/TC 43/SC 1 Noise; IEC/TC 29 Electroacoustics, ISO/TC 108 Mechanical Vibration and Shock; ISO/TC 108/SC 2 Measurement and Evaluation of Mechanical Vibration and Shock as applied to Machines, Vehicles and Structures; ISO/TC 108/SC3 Use and Calibration of Vibration and Shock Measuring Instruments, ISO/TC 108/SC 4 Human Exposure to Mechanical Vibration and Shock; ISO/TC 108/SC 5 Condition Monitoring and Diagnostics of Machines, and ISO/TC 108/SC 6 Vibration and Shock Generating Systems.

• ***Tuesday, 6 June 2006***

Accredited Standards Committee S3 on Bioacoustics.

• ***Tuesday, 6 June 2006***

Accredited Standards Committee S12 on Noise.

• ***Wednesday, 7 June 2006***

Accredited Standards Committee S2 on Mechanical Vibration and Shock.

• ***Wednesday, 7 June 2006***

Accredited Standards Committee S1 on Acoustics.

Standards News from the United States

American National Standards Call for Comment on Proposals Listed

This section solicits comments on proposed new American National Standards and on proposals to revise, reaffirm, or withdrawal approval of existing standards. The dates listed in parenthesis are for information only.

ASA (ASC S1) (Acoustical Society of America)

REVISIONS

BSR S1.40-200x, Sound Calibrators [Revision of ANSI S1.40-1984 (R2001)]

The standard specifies performance requirements for the sound pressure level, frequency, and total distortion generated by a sound calibrator. It also provides requirements for the influence of environmental conditions, for electromagnetic compatibility, and for instrument marking and documentation. It gives details of the tests necessary to verify that a model of a sound calibrator conforms to all the requirements, as well as details of the method for periodic testing of a sound calibrator. (30 January 2006)

ASA (ASC S12) (Acoustical Society of America)

REAFFIRMATIONS

BSR S12.7-1986 (R200x), Measurements of Impulse Noise [Reaffirmation of ANSI S12.7-1986 (R1998)]

Describes methods of measurement of impulse noise and presentation of data. Applies to all kinds of impulse noise, whether discrete event sources or multiple event sources but not to sounds from sources that have specific measurement standards based on the general methods for measurement of quasi-steady noise. Data that may be reported include characteristics of the time variation of the sound pressure (with or without specific frequency weighting) and sound exposure level. (23 January 2006)

AMCA (Air Movement and Control Association)

NEW STANDARDS

BSR/AMCA 301-200x, Methods for Calculating Fan Sound Ratings from Laboratory Test Data (New standard)

This document establishes standard methods for calculating consistent fan sound ratings from laboratory test data. It applies to fans, blowers, exhausters, or other air-moving devices. (30 January 2006)

EIA (Electronic Industries Alliance)

REVISIONS

BSR/EIA 364-28E-200x, Vibration Test Procedure for Electrical Connectors and Sockets (Revision of ANSI/EIA 364-28D-1999)

Details a method to assess the ability of electrical connector components to withstand specified severities of vibration. (14 February 2006)

 0001-4966/2006/119(3)/1289/4/$22.50

ISA (ISA)

NEW STANDARDS

BSR/ISA RP77.60.02-200x, Fossil Fuel Power Plant—Human-Machine Interface Alarms (New standard)

This recommended practice is provided for the benefit of design engineers and, ultimately, plant operators. Use of this recommended practice will result in a more coherent and useful application of plant alarms for operations personnel. The goal of this recommended practice is to reduce alarm discrepancies, clutter, excessive noise levels, and information overload. This recommended practice pertains to alarms displayed using lights or light-boxes, but not software-based alarms (e.g., CRT-based alarms) and is intended to supplement the information provided by ISA-18.1-1979 (R2004), Annunciator Sequences and Specifications. (12 December 2005)

AHAM (Association of Home Appliance Manufacturers)

NEW STANDARDS

BSR/AHAM AC-2-200x, Method for Sound Testing of Portable Household Electric Room Air Cleaners (New standard)

Establishes a method to determine the sound rating of portable household electric room air cleaners. The sound rating is comprised of a set of sound levels that includes (1) A-weighted sound power level and (2) loudness. (19 December 2005)

Initiation of canvasses

The following ANSI-accredited standards developers have announced their intent to conduct a canvass on the proposed American National Standard(s) listed herein in order to develop evidence of consensus for submittal to ANSI for approval as an American National Standard. Directly and materially affected interests wishing to participate as a member of a canvass list, i.e., consensus body, should contact the sponsor of the standard within 30 days of the publication date of this issue of Standards Action. Please also review the section entitled "American National Standards Maintained Under Continuous Maintenance" contained in Standards Action for information with regard to canvass standards maintained under the continuous maintenance option.

AMCA (Air Movement and Control Association)

BSR/AMCA 301-200x, Methods for Calculating Fan Sound Ratings from Laboratory Test Data (New standard)

Erratum

ANSI S1.1-1994 and ANSI S3.20-1995

Accredited Standards Committee S1, Acoustics, and Accredited Standards Committee S3, Bioacoustics, have published an erratum for each of the following documents:

ANSI S1.1-1994 American National Standard Acoustical Terminology and ANSI S3.20-1995 American National Standard Bioacoustical Terminology

Both ANSI S1.1-1994 and ANSI S3.20-1995 contain the following error. In ANSI S1.1-1994 the error is on p. 35, clause 12.06. In ANSI S3.20-1995, it occurs on p. 32, in Annex C, clause C12.06.

The existing text is

12.06 **calculated loudness level**. Loudness level calculated by a specified procedure. Unit, phon.

NOTE 1 Such procedures are given in ANSI S3.4-1980 (R1986), American National Standard Procedure for the Computation of Loudness of Noise, and in ISO 532:1975, Method for Calculating Loudness Level.

NOTE 2 Calculated loudness level in phons is related to loudness in sones by the equation

$$L = 10 \log_2 n_s,$$

where L is loudness level in phons and n_s is loudness in sones.

NOTE 3 A twofold change in loudness corresponds to an interval of 10 phons.

The equation given is incorrect and should be

$$l = 40 + 10 \log_2 n_s,$$

Final Actions on American National Standards

The standards actions listed below have been approved by the ANSI Board of Standards Review (BSR) or by an ANSI-Audited Designator, as applicable.

ASA (ASC S12) (Acoustical Society of America)

REAFFIRMATIONS

ANSI S12.9-Part 6-2000 (R2005), Quantities and Procedures for Description and Measurement of Environmental Sound—Part 6: Methods for Estimation of Awakenings Associated with Aircraft Noise Events Heard in Homes (Reaffirmation of ANSI S12.9-Part 6-2000) (2 December 2005)

REVISIONS

ANSI S12.9-Part 4-2005, Quantities and Procedures for Description and Measurement of Environmental Sound—Part 4: Noise Assessment and Prediction of Long-Term Community Response [Revision of ANSI S12.9-Part 4-1996 (R2001)] (2 December 2005)

Standards News from Abroad

(Partially derived from *ANSI Reporter and ANSI Standards Action*, with appreciation)

Newly Published ISO and IEC Standards

Listed here are new and revised standards recently approved and promulgated by ISO—the International Organization for Standardization

ISO Standards

ACOUSTICS (TC 43)

ISO 389-7:2005, Acoustics—Reference zero for the calibration of audiometric equipment—Part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions.

ISO/DIS 3746, Acoustics—Determination of sound power levels and sound energy levels of noise sources using sound pressure—Survey method using an enveloping measurement surface over a reflecting plane (23 February 2006)

ISO/DIS 10846-5, Acoustics and vibration—Laboratory measurement of vibro-acoustic transfer properties of resilient elements—Part 5: Driving point method for determination of the low frequency transfer stiffness of resilient supports for translatory motion (16 February 2006)

ISO/DIS 10846-1, Acoustics and vibration—Laboratory measurement of vibro-acoustic transfer properties of resilient elements—Part 1: Principles and guidelines (16 February 2006)

ISO/DIS 10846-2, Acoustics and vibration—Laboratory measurement of vibro-acoustic transfer properties of resilient elements—Part 2: Direct method for determination of the dynamic stiffness of resilient supports for translatory motion (16 February 2006)

MECHANICAL VIBRATION AND SHOCK (TC 108)

ISO 18431-1:2005, Mechanical vibration and shock—Signal processing—Part 1: General introduction

TRACTORS AND MACHINERY FOR AGRICULTURE AND FORESTRY (TC 23)

ISO 5008/Cor1:2005, Agricultural wheeled tractors and field machinery—Measurement of whole-body vibration of the operator—Corrigendum

CINEMATOGRAPHY (TC 36)

ISO 22234:2005, Cinematography—Relative and absolute sound pressure levels for motion-picture multi-channel sound systems—Measurement methods and levels applicable to analog photographic film audio, digital photographic film audio and D-cinema audio

ISO Draft Standards

MECHANICAL VIBRATION AND SHOCK (TC 108)

ISO/DIS 21289, Mechanical vibration and shock—Parameters to be specified for the acquisition of vibration data (10 March 2006)

ISO Technical Specifications

HYDROMETRIC DETERMINATIONS (TC 113)

ISO/TS 24154:2005, Hydrometry—Measuring river velocity and discharge with acoustic Doppler profilers.

IEC Standards

ELECTROACOUSTICS (TC 29)

IEC 60118-7 Ed. 2.0 b: 2005, Electroacoustics—Hearing aids—Part 7: Measurement of the performance characteristics of hearing aids for production, supply and delivery quality assurance purposes

IEC 60118-8 Ed. 2.0 b: 2005, Electroacoustics—Hearing aids—Part 8: Methods of measurement of performance characteristics of hearing aids under simulated *in situ* working conditions

IEC Draft Standards

59/434/FDIS, IEC 60704-3 Ed 2.0: Household and similar electrical appliances—Test code for the determination of airborne acoustical noise—Part 3: Procedure for determining and verifying declared noise emission values (6 January 2006)

U.S. TAG	ISO and IEC documents
S1	**Second IEC/CD 60318-6 (29/592/CD)** "Electroacoustics—Simulators of human head and ear—Part 6: Mechanical coupler for the measurements on bone vibrators" (Revision of IEC 60373:1990) (6 January 2006)
S12	**ISO/DIS 3746** "Acoustics—Determination of sound power levels and sound energy levels of noise sources using sound pressure—Survey method using an enveloping measurement surface over a reflecting plane" (3 April 2006)
	ISO/DIS 4869-3 "Acoustics—Hearing protectors—Part 3: Measurement of insertion loss of ear-muff type protectors using an acoustic test fixture" (13 February 2006)
	ISO/DIS 10846-1 "Acoustics and vibration—Laboratory measurement of vibro-acoustic transfer properties of resilient elements— Part 1: Principles and guidelines"
	ISO/DIS 10846-2 "Acoustics and vibration—Laboratory measurement of vibro-acoustic transfer properties of resilient elements— Part 2: Direct method for determination of the dynamic stiffness of resilient supports for translatory motion" (27 March 2006)
	ISO/DIS 10846-5 "Acoustics and vibration—Laboratory measurement of vibro-acoustic transfer properties of resilient elements— Part 5: Driving point method for determination of the low frequency transfer stiffness of resilient supports for translatory motion" (27 March 2006)

Report on International Acoustical Standards Committee Meetings in Canada (George S. K. Wong)

Introduction:

The International meetings involved (1) IEC/TC 29: Electroacoustics (30 May–3 June 2005); (2) ISO/TC 43: Acoustics, ISO/TC 43 SC 1: Noise; ISO/TC 43/SC 2: Building Acoustics (6–10 June 2005); and (3) ISO/TC 108: Mechanical vibration and shock; ISO/TC 108/SC2: Measurement and evaluation of mechanical vibration and shock as applied to machines, vehicles and structures; ISO/TC 108/SC3: Use and calibration of vibration and shock measuring instruments, and ISO/TC 108/SC 6: Vibration and shock generating systems (29 August–2 September 2005). The Canadian Standards Association (CSA) was very kind to provide the excellent CSA building facilities in Mississauga, near Toronto, for the above prestigious venues.

The meetings were hosted by the Standards Council of Canada (SCC) and managed by an organization committee chaired by George Wong of the National Research Council of Canada (NRC).

Meeting topics:

(1) The meetings of IEC/TC 29 consisted of topics on sound level meters, microphones, audiometers, hearing aids and hearing aids induction loops, sound calibrators, and ear simulators. There were 55 delegates from 13 countries. The entertainment during the subsidized banquet was provided by a talented flutist, Ms. Atarah Boyd from France. The delegates also enjoyed a delightful unrehearsed duet made possible by Mr. Hiroaki Takinami of Japan.

(2) The meetings of ISO/TC 43 and its subcommittees consist of topics on threshold of hearing, loudness scaling, hearing protectors, structural-borne sound, machinery noise, transfer properties of resilient elements, sound absorption of road surfaces and pavement surfaces, impulse sound propagation, road vehicle emission noise, environmental noise, shooting range noise, noise from aircraft/civil airports, exposure to noise, sound insulation in buildings, reverberation measurements, new methods, and random-incidence scattering of surfaces. There were 94 delegates from 14 countries. The working group meeting on structure-borne sound was held on a Sunday with the courtesy of Hatch Associates of Mississauga. The entertainment during the subsidized banquet was provided by a talented young Canadian pianist, Mr. Adam Sherkin. Special credit must go to Adam's skill in successfully getting his grand piano into the Chinese restaurant. After his performance there were standing ovations from the delegates.

(3) The meetings of ISO/TC 108 (ASA Secretariat) and its subcommittees consists of topics on terminology, vibration and shock isolation, condition assessment of structural systems from dynamic response measurements, signal processing methods for the analysis of stationary mechanical vibration, vibration materials, measurement of vibration power flow from machines into connected support structures, introduction to balancing, vibration of machines, vibration of ships, vibration of stationary structures, vibration of machines with active magnetic bearings, ground-borne noise and vibration from rail systems, basic techniques for vibration diagnostics, vibration and shock resistance of sensitive equipment, calibration of vibration and shock transducers, vibration condition monitoring transducers and instrumentation, guidance for selection of vibration generators, and shock testing machines. There were 94 delegates from 14 countries. Again, the entertainment during the subsidized banquet was provided by the talented young Canadian pianist, Mr. Adam Sherkin.

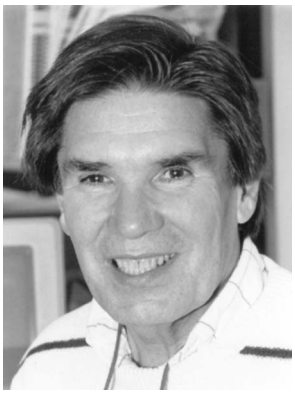
The delegates were very satisfied with the excellent CSA meeting facilities and assistance that included "front-desk general information," technical support for computers, connections to the wireless internet, duplication facilities, and the organization of refreshments during the meetings. The meeting organization committee consisted of David Quirt (NRC), Stephen Keith (Health Canada), Tim Kelsall (Hatch Associates), Chris Krajewsky (Ontario Ministry of Environment), Moustafa Osman (Ontario Power Generation), Ray Coleman (SCC), Holly Bourque (SCC), Lixue Wu (NRC), and George Wong (NRC). The Committee would like to thank CSA for the above very successful international standards meetings.

ACKNOWLEDGMENTS

The Organizing Committee gratefully acknowledges the sponsorship of the following organizations: Acoustical Society of America; ASTM, USA; Brüel and Kjær, Denmark; Canadian Acoustical Association; E-A-R Aearo Co., USA; Etymōtic Research, USA; Eckel Industries of Canada Ltd.; General Motors Corp., USA; G.R.A.S., Denmark; Health Canada; INCE USA; National Research Council Canada; PCB Piezotronics, USA; Quest Technology, USA; Scantek, Inc., USA; and Unitron Industries, Canada.

OBITUARY

William Alfred Watkins·1926–2004



William Alfred Watkins, a pioneer in marine bioacoustics, died on 24 September 2004 at his home in East Falmouth, Massachusetts. He was born in 1926 in Conakry, French Guinea in West Africa to American missionary parents. He spent his youth and adolescence in multilingual West Africa where he acquired fluency in English, French, and Fula, the dominant West African language, and knowledge of many additional African languages. This early training in understanding the nuances of human speech established the ability he was later to employ in deciphering the vocalizations of marine mammals.

William Watkins received his B.A. in Anthropology in 1947 from Wheaton College in Wheaton, Illinois. His interest in human cultures was complemented by an interest in electronics. He worked at the Wheaton College radio station subsequent to graduation and then in 1950 he returned to West Africa to establish a missionary short-wave radio station, ELWA, in Monrovia, Liberia. He returned to the US in 1957. Dr. Watkins received his Ph.D. from the University of Tokyo in 1981. His skills in language acquisition were again evident when he defended his dissertation in Japanese.

In 1958 William Watkins joined the Woods Hole Oceanographic Institution as a research assistant in electronics. His innate interest in natural history combined with his skill in electronics attracted the attention of William E. Schevill, who had published the first recordings of dolphin underwater vocalizations. In order to record other species and begin to learn about the acoustical behavior of cetaceans, much better equipment was required. At the early stage of development of marine mammal bioacoustics, equipment had to be constructed rather than purchased. William Watkins' initial contribution to the collaboration culminated in the "WHOI Rowboat Recorder," to obtain high-frequency recordings from cetaceans at sea.

Dr. Watkins quickly recognized that simply cataloging the vocalizations of marine mammals was only the first step toward understanding the function of these vocalizations in the life of the animals. Although he had little background in natural history or animal behavior, he soon became a keen observer and interpreter. His papers always attempted cautiously to place the recordings in the context of the natural history of the animals. Caution was a hallmark of his work whether it was in terms of approaches to animals, potential disturbance of animals, or interpretation of the data. He was always careful to distinguish between data and interpretation and insisted that the latter never break the bonds with the former.

Both a cautious approach and a deep understanding of electronics, physics, and acoustics placed Dr. Watkins in a unique position in the early development of marine mammal bioacoustics. As more natural historians without this background ventured into marine mammal recording, he taught these fledgling bioacousticians how to interpret spectrograms and avoid simple errors in the interpretation of underwater recordings. Researchers in terrestrial as well as marine bioacoustics have acknowledged his corrective, instructive hand in the analysis of animal sounds.

William Watkins was at the forefront of the development of passive and active acoustic tracking of cetaceans. His four-hydrophone array provided the first understanding of the three-dimensional relationships between vocalizing sperm whales. He was one of the first to track whales using an acoustic transponder and also one of the first to use the Navy's SOSUS hydrophone array to monitor vocalizations and movements across ocean basins of identified whales, and unidentified whales (the subject of a posthumous publication).

Although he is best known to readers of this journal for his acoustic work, his major contributions to another area of marine mammal research must be noted. He developed the first implantable radio tag that could be reliably attached remotely to great whales. This tag was initially used to keep shipboard observers in continual contact with an identified whale whose movements, behaviors, and social interactions could thereby be recorded. Subsequent versions of the tag provided information on diving depth profiles and ultimately developed into a satellite tag. These tags were deployed on fin, humpback, Brydes, and sperm whales in the Caribbean, North Atlantic, and North Pacific.

Over a career of more than 40 years, Dr. Watkins accumulated arguably the most extensive library of marine mammal vocalizations every recorded by one individual: over 20 000 calls from more than 70 species of marine mammals. He recognized that these recordings should be shared with the broader scientific community, and that only a select few individuals could conceptualize the sound of the vocalizations by viewing published sonograms. Thus, his very first marine mammal publication included a phonograph record that set the pattern of extensive sharing of his recordings with the students and fellow researchers. In order to make these recordings more readily available he developed one of the first acoustic databases.

During his career at Woods Hole, William Watkins advanced from Research Assistant to Senior Research Specialist. With his formal retirement in 1996 he became WHOI's first Oceanographer Emeritus. His retirement was only a formality so far as his productivity was concerned. He continued his data acquisition, analysis, interpretation, and publication and his support of his entire research group. He was in the laboratory most days until multiple myeloma severely restricted his activities, but even that debilitation did not deter him from daily contact with his laboratory and continued writing of journal articles, some of which have been published posthumously.

The most fitting tribute to the pioneering work of William A. Watkins is in the bibliography of almost every paper published on marine mammal bioacoustics or marine mammal radio or satellite tracking. Invariably those bibliographies have a number of citations "Watkins, W.A… ." He published close to 200 papers over his distinguished career. His passion was recording the sounds of marine mammals at sea, and more than anyone else he brought the voices of mammals living under the sea to our ears.

He is survived by his wife, Joan Watkins, three sons and two daughters, a brother, a stepson and a stepdaughter, and 12 grandchildren.

D. WARTZOK
P. L. TYACK

 0001-4966/2006/119(3)/1295/1/$22.50

OBITUARY

Francis John Fry·1920–2005

Francis John Fry, a long-time Fellow of the Society, died 22 Sept. 2005.

Frank, as he was known to his colleagues, was born 2 April 1920, in Johnstown, Pa, his mother being an infant survivor of the 1889 Johnstown Flood.

He received the B.S. degree from Pennsylvania State University in 1940 and the M.S. degree from the University of Pittsburgh in 1946, both in electrical engineering. From 1940 to 1946 he was a design engineer for the Westinghouse Electric Corp., where he designed circuitry for the Grand Coulee Dam. This period also included his being a consultant to the Manhattan Project on development of the mass separation techniques for the atomic bomb project

In 1946, Frank joined his brother, William J. (Bill) Fry, at the Department of Electrical Engineering, University of Illinois. Their laboratory, the Bioacoustics Research Laboratory, became internationally renowned for its numerous in-depth and detailed studies of the biological effects of intense ultrasound; the laboratory continues today. The numerous publications of this work, many in JASA, included ultrasonic interferometry, variable frequency SONAR projectors, neurosonic surgery, behavorial studies, anatomic studies, molecular studies, ultrasonic microscopy, and more.

Frank was an exceptional designer of mechanical devices. As an example, he designed a precision stereotoxic appartus and head holder which were used in a 5-year clinical research program at the University of Iowa on the application of high-intensity focused ultrasound to human neurosurgery, a cooperative undertaking between the Universities of Illinois and Iowa. These devices made it possible to position the focal volume of the sound beam precisely at the chosen position in the central nervous system. In the period prior to the human surgical studies, laboratory animals, *viz*., rodents, felines, and primates were studied, for which Frank learned, and carried out, the sterile surgical procedures. Frank's first-hand knowledge of wetlab experimental procedures and his unusual abilities in mechanical design led to instrumentation for human high-intensity ultrasonic surgery that is of great envy today.

In the late 1950s Bill and Frank founded the Interscience Research Institute in Champaign, IL and, among other studies, developed a mechanical artificial heart, on which calves lived for prolonged periods.

In 1972 Frank moved to Indianapolis, where he became the Associate Director of the Indianapolis Center for Advanced Research (ICFAR) and an Associate Professor of Surgery at Indiana University School of Medicine. He also became a development engineer and senior research scientist for Labsonics, Inc. in Mooresville, IN. He broadened his earlier bioultrasonics studies to include the development of new methods for medical tissue visualization.

Throughout his lengthy and productive career Frank published more than 350 journal articles, books, etc. He received approximately 30 patents, and presented more than 80 talks and lectures, domestically and internationally. He was also a Fellow of the American Institute for Ultrasound in Medicine (AIUM), where he received the Pioneer Award for Applications of Ultrasound to Medicine. He also served as chairman of the AIUM Committee on Biological Effects of Ultrasound and as an Associate Editor of the IEEE Transactions on Sonics and Ultrasonics. Frank was also a member of American Society for Artificial Internal Organs, Neurosci. Soc., Biomed. Engr. Soc., and Sigma Xi.

In August 1946 Frank married Marie Ann Michels. They adopted 12 children of different races and nationalities; all were at his bedside when he passed away.

FLOYD DUNN

 0001-4966/2006/119(3)/1296/1/$22.50

OBITUARY

Francis John Fry·1920–2005

Francis John Fry, a long-time Fellow of the Society, died 22 Sept. 2005.

Frank, as he was known to his colleagues, was born 2 April 1920, in Johnstown, Pa, his mother being an infant survivor of the 1889 Johnstown Flood.

He received the B.S. degree from Pennsylvania State University in 1940 and the M.S. degree from the University of Pittsburgh in 1946, both in electrical engineering. From 1940 to 1946 he was a design engineer for the Westinghouse Electric Corp., where he designed circuitry for the Grand Coulee Dam. This period also included his being a consultant to the Manhattan Project on development of the mass separation techniques for the atomic bomb project

In 1946, Frank joined his brother, William J. (Bill) Fry, at the Department of Electrical Engineering, University of Illinois. Their laboratory, the Bioacoustics Research Laboratory, became internationally renowned for its numerous in-depth and detailed studies of the biological effects of intense ultrasound; the laboratory continues today. The numerous publications of this work, many in JASA, included ultrasonic interferometry, variable frequency SONAR projectors, neurosonic surgery, behavorial studies, anatomic studies, molecular studies, ultrasonic microscopy, and more.

Frank was an exceptional designer of mechanical devices. As an example, he designed a precision stereotoxic appartus and head holder which were used in a 5-year clinical research program at the University of Iowa on the application of high-intensity focused ultrasound to human neurosurgery, a cooperative undertaking between the Universities of Illinois and Iowa. These devices made it possible to position the focal volume of the sound beam precisely at the chosen position in the central nervous system. In the period prior to the human surgical studies, laboratory animals, *viz*., rodents, felines, and primates were studied, for which Frank learned, and carried out, the sterile surgical procedures. Frank's first-hand knowledge of wetlab experimental procedures and his unusual abilities in mechanical design led to instrumentation for human high-intensity ultrasonic surgery that is of great envy today.

In the late 1950s Bill and Frank founded the Interscience Research Institute in Champaign, IL and, among other studies, developed a mechanical artificial heart, on which calves lived for prolonged periods.

In 1972 Frank moved to Indianapolis, where he became the Associate Director of the Indianapolis Center for Advanced Research (ICFAR) and an Associate Professor of Surgery at Indiana University School of Medicine. He also became a development engineer and senior research scientist for Labsonics, Inc. in Mooresville, IN. He broadened his earlier bioultrasonics studies to include the development of new methods for medical tissue visualization.

Throughout his lengthy and productive career Frank published more than 350 journal articles, books, etc. He received approximately 30 patents, and presented more than 80 talks and lectures, domestically and internationally. He was also a Fellow of the American Institute for Ultrasound in Medicine (AIUM), where he received the Pioneer Award for Applications of Ultrasound to Medicine. He also served as chairman of the AIUM Committee on Biological Effects of Ultrasound and as an Associate Editor of the IEEE Transactions on Sonics and Ultrasonics. Frank was also a member of American Society for Artificial Internal Organs, Neurosci. Soc., Biomed. Engr. Soc., and Sigma Xi.

In August 1946 Frank married Marie Ann Michels. They adopted 12 children of different races and nationalities; all were at his bedside when he passed away.

FLOYD DUNN

 0001-4966/2006/119(3)/1296/1/$22.50

REVIEWS OF ACOUSTICAL PATENTS

Lloyd Rice
11222 Flatiron Drive, Lafayette, Colorado 80026

The purpose of these acoustical patent reviews is to provide enough information for a Journal reader to decide whether to seek more information from the patent itself. Any opinions expressed here are those of reviewers as individuals and are not legal opinions. Printed copies of United States Patents may be ordered at $3.00 each from the Commissioner of Patents and Trademarks, Washington, DC 20231. Patents are available via the Internet at http://www.uspto.gov.

Reviewers for this issue:

GEORGE L. AUGSPURGER, *Perception, Incorporated, Box 39536, Los Angeles, California 90039*
JOHN M. EARGLE, *JME Consulting Corporation, 7034 Macapa Drive, Los Angeles, California 90068*
JOHN ERDREICH, *Ostergaard Acoustical Associates, 200 Executive Drive, West Orange, New Jersey 07052*
SEAN A. FULOP, *California State University, Fresno, 5245 N. Backer Avenue M/S PB92, Fresno, California 93740-8001*
JEROME A. HELFFRICH, *Southwest Research Institute, San Antonio, Texas 78228*
DAVID PREVES, *Starkey Laboratories, 6600 Washington Ave. S., Eden Prarie, Minnesota 55344*
DANIEL R. RAICHEL, *2727 Moore Lane, Fort Collins, Colorado 80526*
CARL J. ROSENBERG, *Acentech Incorporated, 33 Moulton Street, Cambridge, Massachusetts 02138*
NEIL A. SHAW, *Menlo Scientific Acoustics, Inc., Post Office Box 1610, Topanga, California 90290*
ERIC E. UNGAR, *Acentech, Incorporated, 33 Moulton Street, Cambridge, Massachusetts 02138*
ROBERT C. WAAG, *University of Rochester, Department of Electrical and Computer Engineering, Rochester, New York 14627*

6,949,869

43.35.Zc PIEZOELECTRIC ACTUATOR, LIQUID JETTING HEAD INCORPORATING THE SAME, PIEZOELECTRIC ELEMENT, AND METHOD OF MANUFACTURING THE SAME

Chang Junhua *et al.*, assignors to Seiko Epson Corporation
27 September 2005 (Class 310/328); filed in Japan 18 March 2002

This patent describes a piezoelectric ink-jet driver with features for stiffening the driven region for high-frequency operation. The concept is simple but the description is very difficult to follow. There appears to be nothing novel here, but claims are made for all sorts of construction details. Not much can be learned from reading this.—JAH

6,952,073

43.35.Zc VIBRATION WAVE DRIVING APPARATUS, VIBRATION MEMBER AND DRIVING SYSTEM FOR THE VIBRATION WAVE DRIVING APPARATUS

Shinji Yamamoto, assignor to Canon Kabushiki Kaisha
4 October 2005 (Class 310/323.06); filed in Japan 18 April 2002

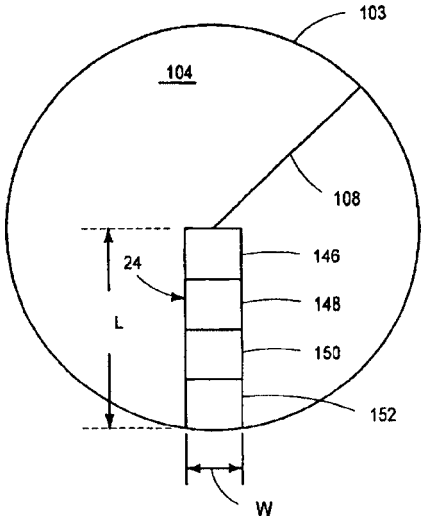
A vibration-wave-driven motor is described which requires only two phases of electrical drive signal to its piezoelectric rotor. The simplification of the electrical drive is obvious, but the cost in torque and cogging performance is not mentioned. The patent is brief and concise, but lacking in supporting data or theory.—JAH

6,946,774

43.35.Zc SEGMENTED UNIFORM ENERGY MEGASONIC TRANSDUCER

Mark J. Beck *et al.*, assignors to Product Systems Incorporated
20 September 2005 (Class 310/317); filed 13 September 2004

It is not often that one sees the term "megasonic" used. That may be the reason why this patent was written, because there is nothing disclosed here that the average person would not find obvious. The inventors have come up with a way of irradiating a spinning disk (a silicon wafer or perhaps a disc drive platter) with ultrasound in inverse proportion to the length



of time it spends under the beam as it spins. Actually, there are three different techniques presented, all of which work somewhat like that shown in the figure. Transducers **146**, **148**, **150**, **152** form a shaded array that treats each annular region of the substrate with a power proportional to its linear velocity. This all makes sense, but is it really worth the trouble?—JAH

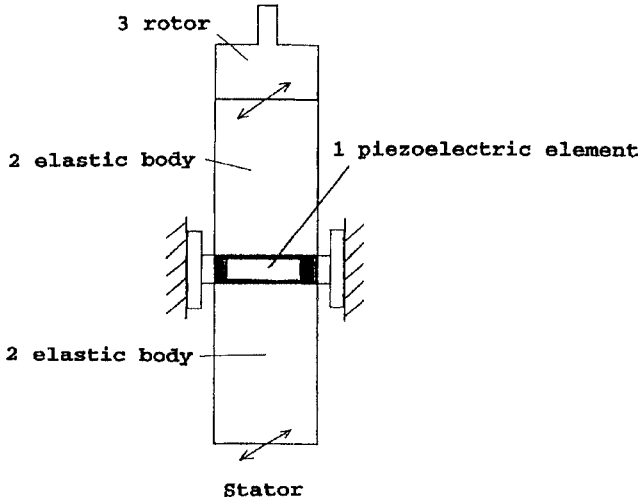
6,952,072

43.35.Zc ULTRASONIC MOTOR AND ELECTRONIC APPARATUS UTILIZING ULTRASONIC MOTOR

Akihiro Iino *et al.*, assignors to Seiko Instruments Incorporated
4 October 2005 (Class 310/323.04); filed in Japan 9 November 2001

This patent describes an ultrasonic motor for application to drilling. Referring to the figure, piezoelectric disk **1** oscillates in a thickness mode, driving helically grooved bodies **2**. These bodies engage projections in the

 0001-4966/2006/119(3)/1299/15/$22.50



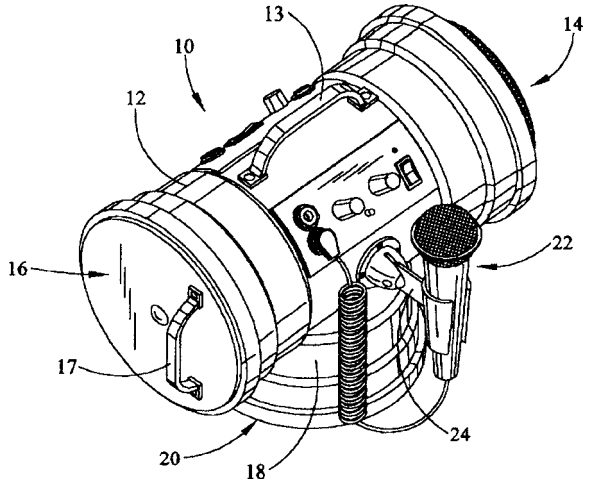
housing of the device so as to convert their longitudinal motion into rotation. The upper body comes into intermittent contact with rotor **3**, sending it into unidirectional rotation. The patent is not very clearly written, but it does show some interesting alternative configurations.—JAH

6,947,566

43.38.Ar METHOD AND APPARATUS FOR A PORTABLE PUBLIC ADDRESS SYSTEM

Raymond L. Mears, Calhoun, Louisiana
20 September 2005 (Class 381/75); filed 2 October 2001

The battery-powered ''loud hailer'' has been with us for many decades, and this *omnium gatherum* patent offers a long list of embellishments to the original design. In addition to its basic functions, ''improved options include



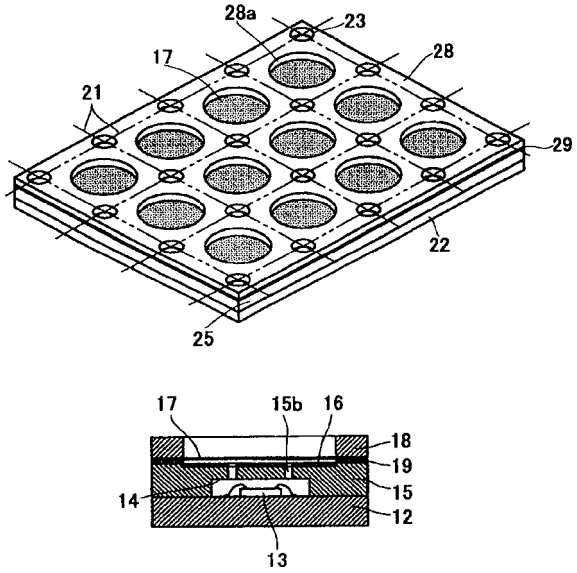
a unique compact configuration that allows for modular construction of all components for plug and play operations and rapid field replacement; a weatherproof housing for operation in inclement weather, and a tape player to provide prerecorded programming.''—JME

6,947,568

43.38.Ar CONDENSER MICROPHONE

Haruhisa Tanabe and Megumi Horiuchi, assignors to Citizen Electronics Company, Limited
20 September 2005 (Class 381/174); filed in Japan 15 May 2001

The patent deals with improvements in microphone fabrication via



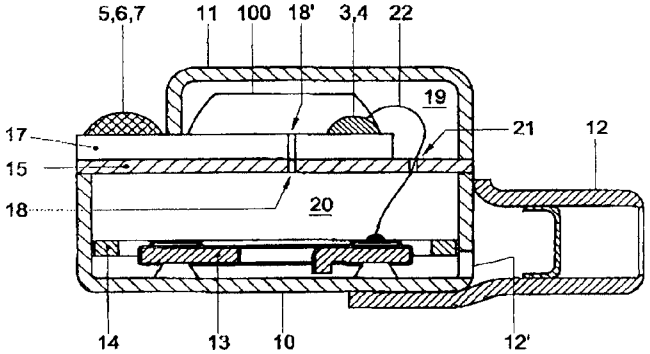
micro-electro-mechanical-systems (MEMS) techniques. It is concise and well written.—JME

6,950,529

43.38.Ar SYSTEM CONSISTING OF A MICROPHONE AND AN AMPLIFIER

Engbert Wilmink, assignor to Sonionmicrotronic Nederland B.V.
27 September 2005 (Class 381/174); filed in the Netherlands 17 May 2000

The patent describes the fairly basic use of lumped elements in the



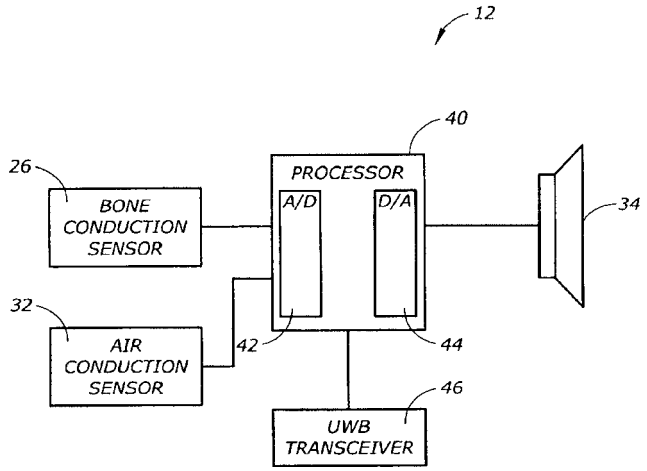
mechanical circuit of the microphone to adjust the overall equalization of a microphone-amplifier system.—JME

6,952,483

43.38.Ar VOICE TRANSMISSION APPARATUS WITH UWB

Peter V. Boesen and James F. Boesen, Jr., assignors to Genisus Systems, Incorporated
4 October 2005 (Class 381/326); filed 5 September 2002

The patent deals with two-way communication in high-noise environments. A multielement sensor is positioned at, and in, the ear canal to cover microphone requirements and those of bone and air conduction for



receiving. The application of ultra-wide-band (UWB) transmission to and from a body pack worn by the user is included in the patent description, presumably to keep the ear assembly as small as possible.—JME

6,946,763

43.38.Dv VIBRATION MOTOR

Young Il Park, assignor to LG Innotek Company, Limited
20 September 2005 (Class 310/81); filed in the Republic of Korea 29 April 2004

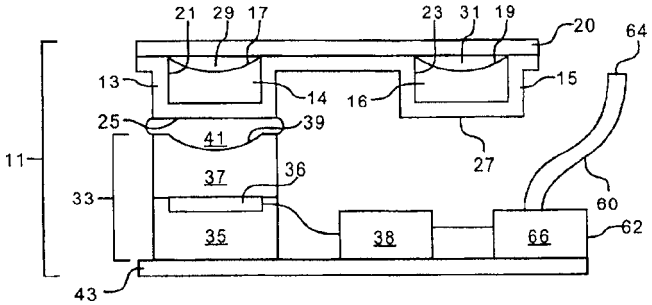
A simple motor patent gives rise to an entertaining read: Obsessed with efficiency, the inventors try to juggle the positions of the contacts on an electrodynamic motor to reduce torque ripple from 38% to 35%. This is really a case of diminishing returns of questionable value, but the inventors seem not to worry.—JAH

6,932,097

43.38.Fx ACOUSTIC CONTROL OF THE COMPOSITION AND/OR VOLUME OF FLUID IN A RESERVOIR

Richard N. Ellson and Mitchell W. Mutz, assignors to Picoliter Incorporated
23 August 2005 (Class 137/2); filed 18 June 2002

According to this patent, devices are needed in the pharmaceutical and biotechnology industries for handling very small amounts of fluids. The patented device would use acoustical energy to collect, mix, and dispense multiple fluids in amounts on the order of a few molecules. The figure shows the device set to monitor the composition or volume within reservoir **13**. An



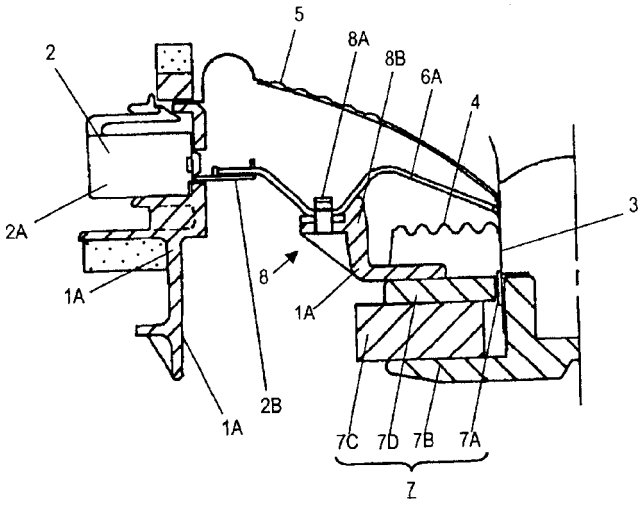
acoustic wave or pulse, produced by generator **35**, is transmitted into focusing device **37**, and then coupled by medium **41** into the reservoir. An echo may be picked up by piezoelectic transducer **36**. By removing cover **20**, a droplet may be ejected and deposited on a nearby substrate (not shown here). Fluid source **64** may be brought over to the uncovered reservoir.—DLR

6,922,477

43.38.Ja SPEAKER

Junzo Ikeyama, assignor to Matsushita Electric Industrial Company, Limited
26 July 2005 (Class 381/409); filed 2 October 2000

The lead wires **6A** from the connection **2** on the speaker frame **1A** to the voice coil **3** have many ways of failing. One method of installing these wires is to place them between the speaker cone and the centering spider. The problem then is how to secure the wires so that they do not touch the cone or spider when the speaker is in use. This patent describes what is said



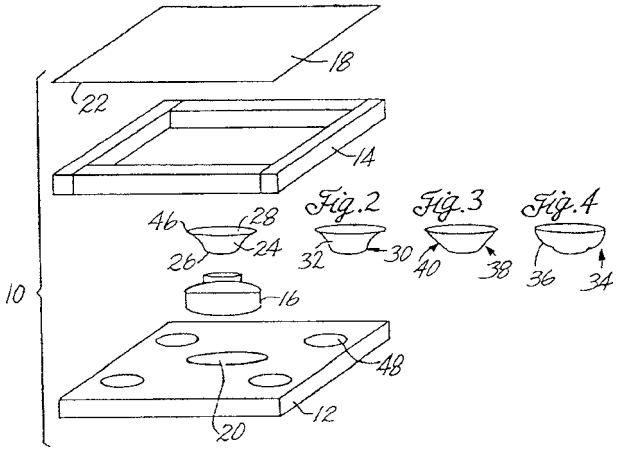
to be an improved intermediate post **8** between the frame contacts **2B** and the voice coil. The design of the post clamping section **8A** is said to facilitate the installation of the lead wires during manufacturing as well as to reduce the number of components for this piece compared to prior art.—NAS

6,925,191

43.38.Ja FLAT PANEL SPEAKER

Michael Petroff *et al.*, assignors to Digital Sonics LLC
2 August 2005 (Class 381/423); filed 23 January 2002

An inexpensive electrodynamic driver **16** is connected to a thin film membrane **18** using an enhancer **24** that has a neck **26** and mouth **28**. The back plate **12** contains a space **20** for the driver and a number of sound breathers **48**. The patent asserts that the frequency response and other pa-



rameters can be modified by changing the shape of the enhancer, the shape and location of the space for the driver, as well as the driver itself, the thickness of the film, the material for the film, perforating the film, and

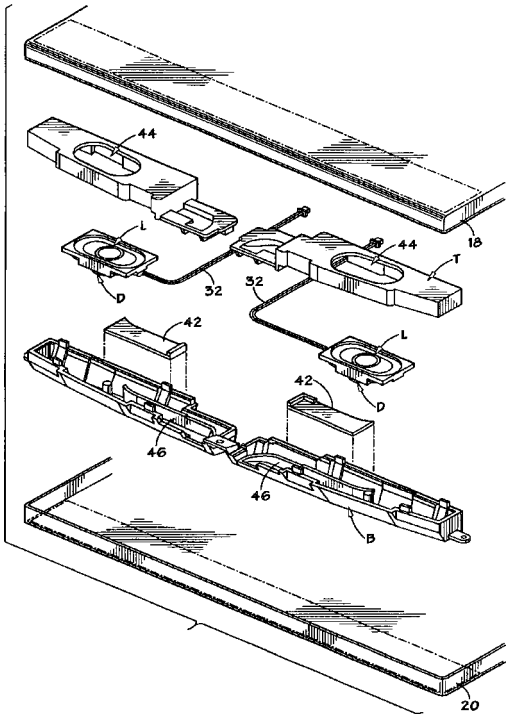
using a magnetic oil in the driver (which sounds like prior art), among other things.—NAS

6,925,188

43.38.Ja PORTED SPEAKER ENCLOSURE OF A PORTABLE COMPUTER

Mitch A. Markow *et al.*, assignors to Hewlett-Packard Development Company, L.P.
2 August 2005 (Class 381/306); filed 20 June 1997

The loudspeakers used in portable laptop computers are by necessity quite small and, in most cases, of limited fidelity. The space in these computers is also at a premium. When an extra octave in response at the low end is deemed desirable and the space is available, a means is described of using a ported enclosure comprising top part **T**, bottom part **B**, port covers **42**, and loudspeakers **L** (also referred to as drivers **D** in another part of the patent).



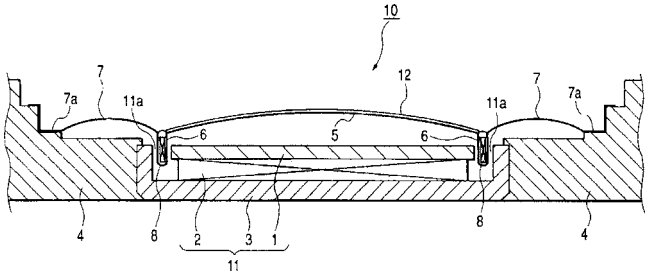
A short, well-written discussion is included in the body of the patent of the factors limiting low-frequency reproduction in small loudspeakers and why ported enclosures can be used to extend low-frequency response.—NAS

6,944,310

43.38.Ja SPEAKER APPARATUS

Masanori Ito *et al.*, assignors to Pioneer Corporation
13 September 2005 (Class 381/386); filed in Japan 4 March 2002

For practical reasons, tiny loudspeakers are usually designed as edge-driven domes. If the diaphragm is not circular, but oval or a rounded rectangle, then the voice coil must also be formed to a noncircular shape, leading to unwanted complexity in the manufacturing process. This patent



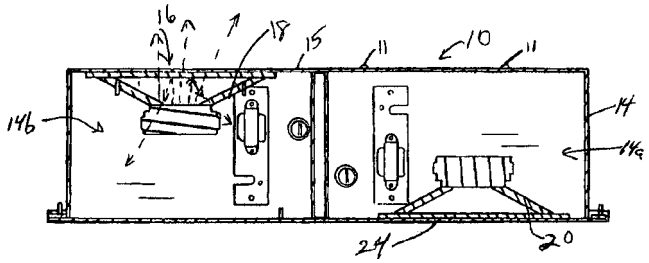
describes a diaphragm that includes a little slot **8** in which voice coil **6** is securely held. Internal stresses cannot deform the coil, and the entire assembly can be centered with a single jig.—GLA

6,950,526

43.38.Ja SOUND MASKING AND PAGING SYSTEM

Jeffrey R. Garstick, assignor to Lowell Manufacturing Company
27 September 2005 (Class 381/73.1); filed 22 April 2003

A conventional downward-firing paging speaker and an upward-firing



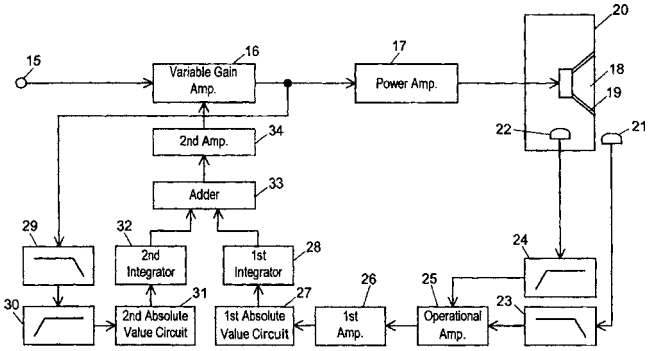
noise masking speaker are housed in a single assembly that replaces a lay-in ceiling tile.—GLA

6,944,302

43.38.Lc SOUND REPRODUCING SYSTEM

Masahide Onishi *et al.*, assignors to Matsushita Electric Industrial Company, Limited
13 September 2005 (Class 381/59); filed in Japan 19 June 2001

For more than 50 years, audio engineers have struggled with the problem of automatically adjusting paging system volume as background noise varies. Today, Japanese electronics manufacturers seem determined to design automotive sound systems that exactly compensate for background noise masking. In both applications, the trick is to somehow measure noise while ignoring the signal from the loudspeaker itself. Some prior art places a sensing microphone very close to the loudspeaker and then compares its output with the electronic signal that drives the loudspeaker. The scheme



described in this patent uses two sensing microphones—one in front of the speaker and one in the back cavity. If the two signals are suitably scaled, then, in the absence of noise, their outputs combine to a simple bandpass function that can be compared to a corresponding electronic bandpass function. Any imbalance must be due to noise, causing amplifier gain to shift

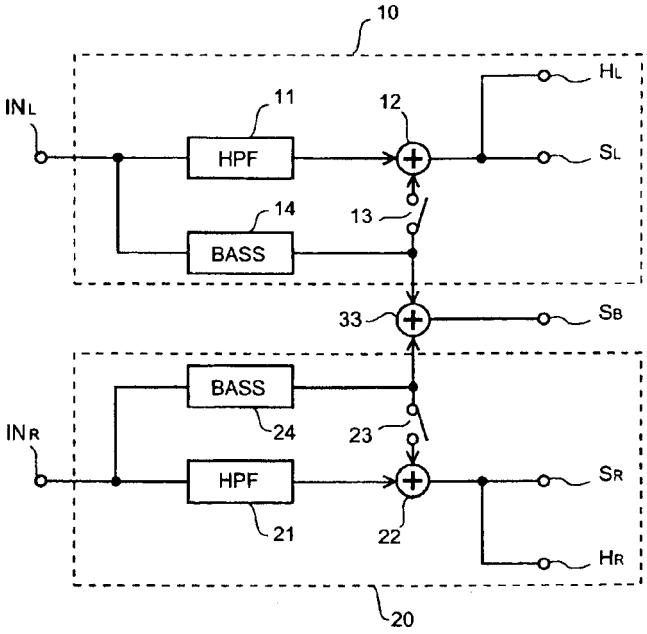
until a null is again achieved. Theoretically, this all makes sense if the loudspeaker operates as a perfect piston mounted in a small, rigid, airtight enclosure.—GLA

6,920,224

43.38.Lc DEEP BASS SOUND BOOSTER DEVICE

Takashi Oki, assignor to Rohm Company, Limited
19 July 2005 (Class 381/123); filed in Japan 20 October 2000

A means of changing the response compensation (equalization) of a two-channel signal is described. Input signals are sent through highpass filters **11**, **21** with output at a relative level of 0 dB and low-frequency filters **14**, **24** with more gain in their pass bands. When headphones are connected



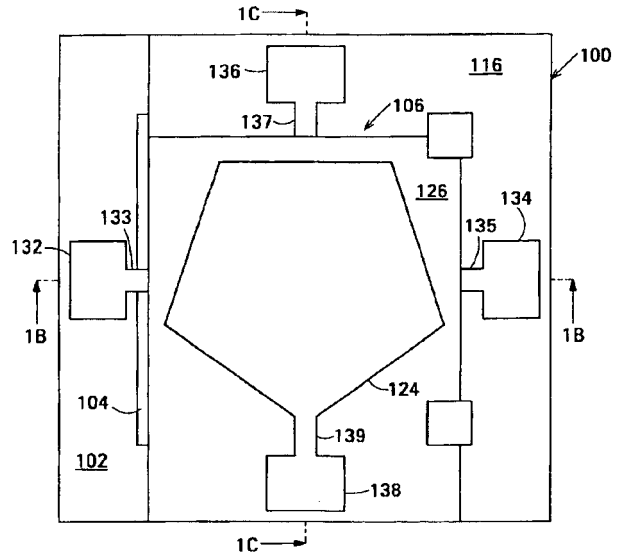
to the device with the bass filters active, the signals from the bass filters are added via **12** and **22** to the outputs, whereas when loudspeakers are connected, the subbass signal is sent to a separate subbass loudspeaker S_B.—NAS

6,946,928

43.38.Rh THIN-FILM ACOUSTICALLY-COUPLED TRANSFORMER

John D. Larson III and Richard C. Ruby, assignors to Agilent Technologies, Incorporated
20 September 2005 (Class 333/189); filed 30 October 2003

This patent describes an acoustical transformer for use in the 2-GHz portion of the rf spectrum. The figures show how the two transformers are constructed of two FBAR (thickness mode) resonators **124**, **126**, and **122**. The resonators are stacked vertically and acoustically coupled by layer **130**, an insulator of low (acoustical) impedance. The effect is to weakly couple the two resonant structures and broaden the frequency range of the transformer coupling. Some guidance is given on how to select the coupling layer



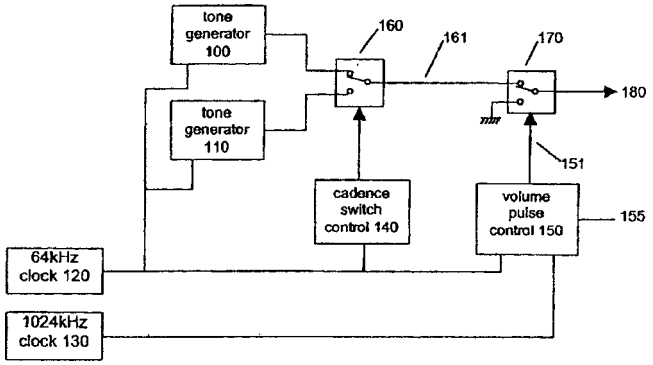
material and thickness to optimize the transformer response for given electrode and resonator materials. There is a great deal of attention paid to balancing the electrical impedances to ground of the primary and secondary structures. This is a detailed and well-written patent.—JAH

6,944,284

43.38.Si PULSED VOLUME CONTROL OF A MAGNETIC RINGER

Chris Goodings, assignor to VTech Communications, Limited
13 September 2005 (Class 379/373.02); filed in the United Kingdom 29 December 2000

There was a time when a telephone's ring tone was generated by a bell, and its loudness was adjusted mechanically. Today, a small transducer is driven by an electronic tone generator and the volume is usually adjusted by a potentiometer. Such an arrangement involves both analog and digital circuitry and could be simplified by keeping everything in the digital domain.



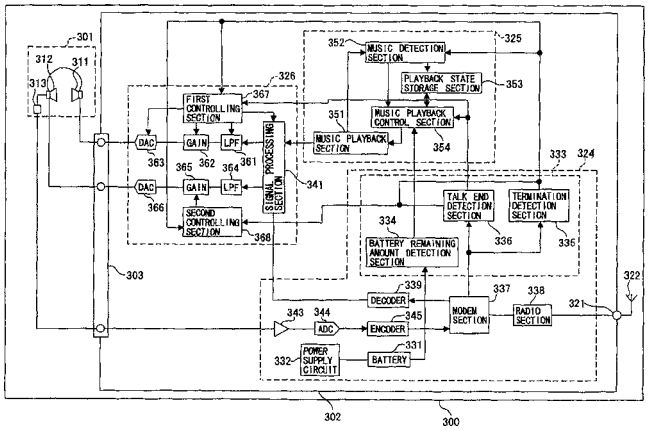
Well, if the driving signal is a square wave, then varying the pulse width will vary the volume level. This technique is used in some telephones. The problem is that tonal quality also changes. This patent describes a constant-timbre driving circuit in which a fixed-amplitude tone is amplitude modulated by a higher-frequency, digital pulse train.—GLA

6,944,287

43.38.Si PORTABLE TERMINAL

Ukyo Mori, assignor to NEC Corporation
13 September 2005 (Class 379/390.01); filed in Japan 18 May 2000

Manufacturers seem eager to produce ''handyphones''—devices that combine stereo music playback with all the functions of an up-to-date cell phone. In designing such a product, numerous decisions must be made concerning interrelations between various functions: Should the ring signal switch automatically to earphones when music is being played? Should cell



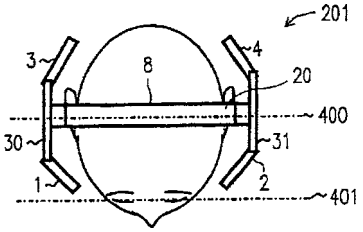
phone voice and warning signals be reproduced through one or both earphones? At the termination of a conversation, should music resume at the point where it was interrupted? Should the user be allowed to listen to music if the battery is low? This patent takes up all the permutations and combinations in considerable, and sometimes confusing, detail. Flow charts and circuit diagrams are used to describe an improved control program.—GLA

6,944,309

43.38.Si HEADPHONE SYSTEM

Kenichi Terai *et al.*, assignors to Matsushita Electric Industrial Company, Limited
13 September 2005 (Class 381/309); filed in Japan 2 February 2000

A number of circuits have been developed to convert multichannel surround sound into two-channel sound for headphone listening. In such systems, it is difficult to encode front-rear information convincingly because auditory front-rear discrimination relies on head-related transfer functions,



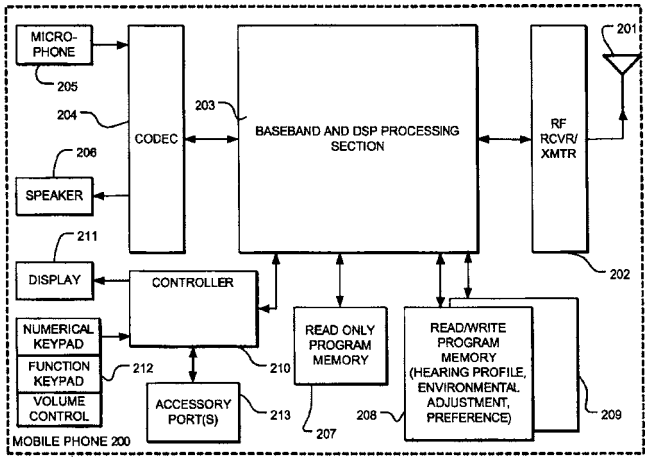
and these vary from one person to the next. This patent describes a system in which four ''nearphones'' are carefully located with respect to the listener's ears. Fairly complicated matrix circuitry is still needed, but the resulting sonic images are said to be robust for almost all listeners.—GLA

6,944,474

43.38.Si SOUND ENHANCEMENT FOR MOBILE PHONES AND OTHER PRODUCTS PRODUCING PERSONALIZED AUDIO FOR USERS

R. Scott Rader *et al.*, assignors to Sound ID
13 September 2005 (Class 455/550.1); filed 20 September 2001

Self-adjusting audio and video systems are available that can automatically compensate for room acoustics and illumination. Some automobiles feature selectable presets for the driver's seat so that each person who drives that car can call up appropriate seat adjustments. This patent describes a



personalized cell phone that not only stores user preferences but can test the user's hearing and apply appropriate electronic compensation. Numerous other features are available, largely achieved through programmable multiband compression.—GLA

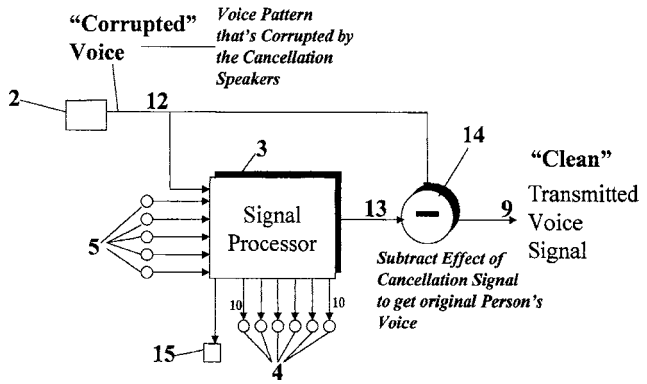
6,952,474

43.38.Si PHONE PRIVACY AND UNOBTRUSIVENESS VIA VOICE CANCELLATION

Edward R. Wittke and Melissa A. Wittke, both of Warwick, New York
4 October 2005 (Class 379/406.02); filed 15 March 2001

Cell phone users would like to maintain privacy while conversing in public areas. One patented solution places a small voice-canceling loudspeaker near the user's mouth, but far enough from the telephone receiver to allow usable pickup. Considering all of the variables involved, practical operation seems iffy at best. Moreover, the patent at hand argues that such a

Figure 3 Subtract out Effects of Cancellation Signal



scheme distorts the acoustic signal picked up by the microphone and transmitted to the other party. The patent teaches that ''clean'' voice transmission can be achieved by somehow electronically subtracting the effect of the cancellation signal. ''Today, technology exists which makes this novel concept feasible and marketable.'' End of discussion. In a more elaborate em-

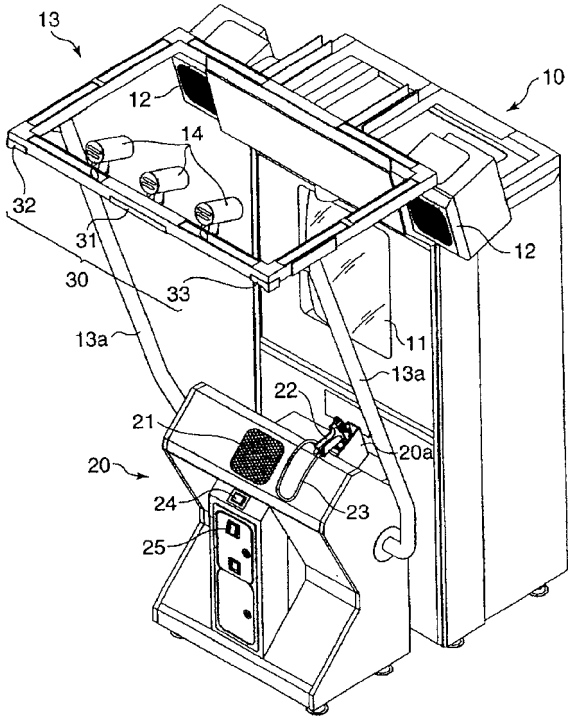
bodiment, the user is surrounded by arrays of microphones and loudspeakers to prevent sound from exiting or entering a personal privacy zone. For those uncomfortable with such cutting edge technology, a plexiglass helmet may be the answer.—GLA

6,918,829

43.38.Tj FIGHTING VIDEO GAME MACHINE

Masahiro Ikariko, assignor to Konami Corporation
19 July 2005 (Class 463/5); filed in Japan 11 August 2000

First-person shooter video games are very popular—in arcades, on the personal computer, and with the several video game boxes now on the market. The patent describes a first-person shooter arcade game that uses signal processing to determine where shooting from the computer side



comes from, whether they ''hit'' the gamer, position of the gamer and the computer antagonist, center of gravity, left/right foot positions, etc., to calculate the proper sound effect to be presented to loudspeakers **12** and **21** to improve the ''realism'' of the game experience.—NAS

6,943,484

43.40.Cw TUNABLE PIEZOELECTRIC MICRO-MECHANICAL RESONATOR

William W. Clark and Qing-Ming Wang, assignors to University of Pittsburgh
13 September 2005 (Class 310/334); filed 6 December 2002

A tunable cantilever resonator is described as operating in the low MHz range. The cantilever has both a sensor and an actuator element on it, and the sensor element is used with various capacitive shunts to obtain a small ($\sim$3%) degree of tuning. The patent asserts that this is a high-Q design, but no evidence of this is provided. A brief discussion is given of how the tuning range can be predicted for a given piezo configuration.—JAH

6,943,647

43.40.Sk BULK ACOUSTIC WAVE FILTER WITH A ROUGHENED SUBSTRATE BOTTOM SURFACE AND METHOD OF FABRICATING SAME

Robert Aigner *et al.*, assignors to Infineon Technologies AG
13 September 2005 (Class 333/187); filed 6 May 2004

This patent describes a bulk acoustic resonator that purports to reinvent the concept of a beam dump. The concept is that creating a diffuse reflective surface beyond the back side of a bulk acoustic resonator has utility in reducing the ripples in the frequency response of such a device. This may be true, but it really is not enough to scatter the acoustic energy. It also needs to be absorbed. The patent is very brief and tells little about the heuristics for locating such a nonreflective surface, but it does contain data showing rough versus smooth backing response from a case study in the 980-MHz rf band.—JAH

6,943,648

43.40.Sk METHODS FOR FORMING A FREQUENCY BULK ACOUSTIC RESONATOR WITH UNIFORM FREQUENCY UTILIZING MULTIPLE TRIMMING LAYERS AND STRUCTURES FORMED THEREBY

Jose Maiz *et al.*, assignors to Intel Corporation
13 September 2005 (Class 333/188); filed 1 May 2003

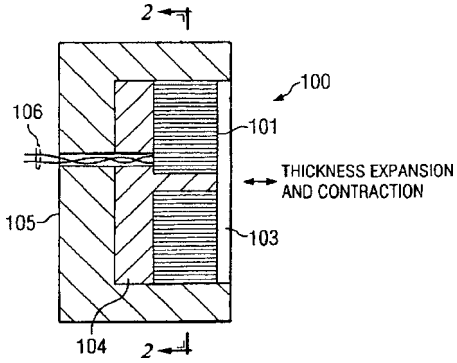
This is your basic thickness-mode resonator, augmented by mass loading. Changing the load allows one to trim the resonant frequency after the resonator has been fabricated. The methods offered are somewhat more suited for MEMS application because of the reliance on removal of mass in etching and laser ablation processes not commonly found in other labs.—JAH

6,946,777

43.40.Sk POLYMER FILM COMPOSITE TRANSDUCER

Thomas E. Owen, assignor to Southwest Research Institute
20 September 2005 (Class 310/334); filed 8 January 2004

This patent discloses a composite PVDF/silicone transducer suitable for efficient coupling to gases. The piezoelectric material is used in the 3-1



mode and can be easily fashioned into a jellyroll for leading and packaging as shown in the figure. The discussion is clear and concise, and some construction tips are given for transducers in the 30–400-kHz range.—JAH

6,942,586

43.40.Tm VIBRATION DAMPENING MATERIAL

Thomas Falone *et al.*, assignors to Sting Free Technologies Company
13 September 2005 (Class 473/520); filed 28 May 2004

This vibration damping material provides a sting-free grip for the handle of athletic equipment (like a golf club), tools, or handlebars of a motorized vehicle. The material does so by introducing a high-tensile fibrous material between elastomeric layers of vibration-absorbing material.—CJR

6,943,290

43.40.Tm VIBRATION DAMPER FOR OVERHEAD POWER LINES

Philip Wellesley Dulhunty, assignor to Dulhunty Power (Aust) Pty. Limited
13 September 2005 (Class 174/42); filed in Australia 14 July 2000

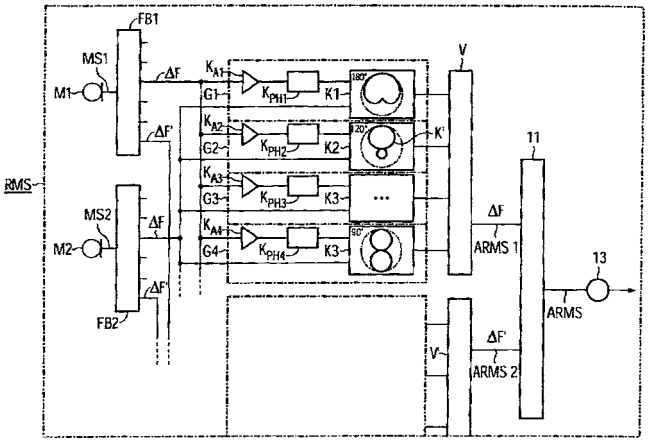
This patent pertains to an improved Stockbridge damper. Such a damper typically consists of a pair of weights attached to the ends of a short length of steel cable, which, at its mid-point, is attached to the power line whose vibrations are to be attenuated. The weights and cable are selected so that the system's natural frequency matches that of the power line. The present patent describes connections between the weights and the cable that do not clamp all of the strands of the cable, so that sliding of the strands relative to each other can increase the system's energy dissipation capability.—EEU

6,950,528

43.60.Fg METHOD AND APPARATUS FOR SUPPRESSING AN ACOUSTIC INTERFERENCE SIGNAL IN AN INCOMING AUDIO SIGNAL

Eghart Fischer, assignor to Siemens Audiologische Technik GmbH
27 September 2005 (Class 381/92); filed in Germany 25 March 2003

The patent deals with microphone arrays in hearing aid applications. Normally, the desired signal originates directly ahead of the user, and the



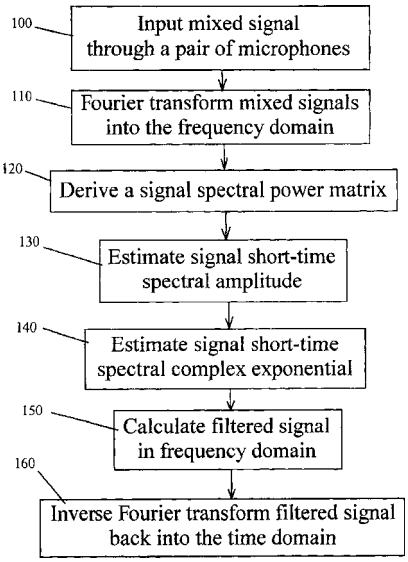
normalized outputs of a number of individual directional microphones (two or three at each ear) are compared. The microphones with the lowest interference components are then selected and processed.—JME

6,952,482

43.60.Fg METHOD AND APPARATUS FOR NOISE FILTERING

Radu Victor Balan and Justinian Rosca, assignors to Siemens Corporation Research, Incorporated
4 October 2005 (Class 381/94.1); filed 5 December 2001

The patent describes transform methods for reducing noise in a communications system. "One known approach to this problem is a two-microphone system, wherein two microphones are placed at fixed locations within the room or vehicle and are connected to a signal processing device.



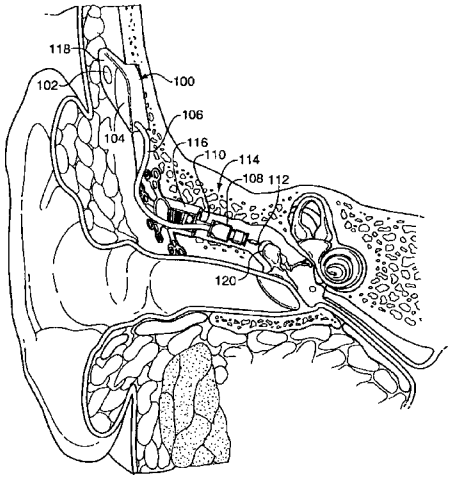
The speaker is assumed to be static during the entire use of this device. The goal is to enhance the target signal by filtering out noise based on the two-channel recording with two microphones." The mathematical bases for the operations are thoroughly outlined, as shown in the figure.—JME

6,945,999

43.66.Ts IMPLANTABLE HEARING AID TRANSDUCER WITH ACTUATOR INTERFACE

Robert Edwin Schneider and Scott Allan Miller III, assignors to Otologics LLC
20 September 2005 (Class 623/10); filed 27 January 2003

An actuator interface **112** that is reshapeable *in situ* is said to continu-



ally ensure good coupling from the implanted transducer **108** to the ossicles

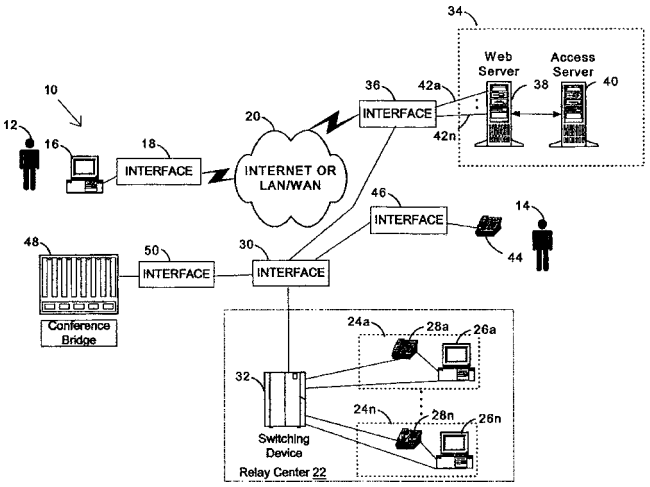
120. This feature is said to mitigate movements due to tissue growth or other changes in loading on the actuator.—DAP

6,950,500

43.66.Ts INTERNET-BASED AND NETWORK-BASED RELAY CENTER ACCESS FOR THE HEARING AND SPEECH IMPAIRED

Pawan Chaturvedi *et al.*, assignors to Sprint Communications Company L.P.
27 September 2005 (Class 379/52); filed 28 November 2000

This is a system for providing relay-center-based translation of telephone messages from text-to-speech and vice versa, as required. The following advantages are cited for this system: (1) a hearing-impaired user's dependence on TDD is eliminated, (2) greater mobility is provided for the



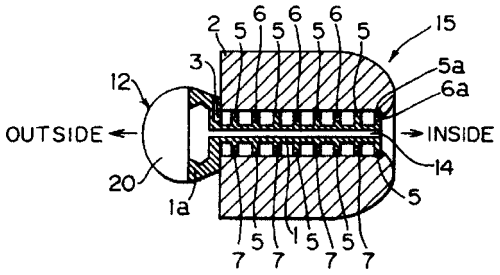
user, (3) the system allows for more relay-center features, (4) multiparty conversations are rendered possible, and (5) full-duplex protocol sessions are enabled with a communications assistant, wherein each party can send and receive data simultaneously.—DRR

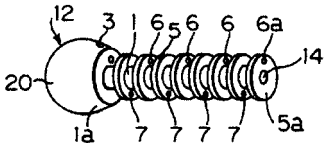
6,761,173

43.66.Vt EAR PLUG TO BE INSERTED INTO THE EXTERNAL AUDITORY CANAL

Takashi Kuno *et al.*, assignors to Mimy Electronics Company, Limited
13 July 2004 (Class 128/864); filed 7 January 2003

This is an earplug useful as a hearing protector or for a hearing aid. Pressure in the external canal is equalized with atmosphere through small





holes **6**, **7** in the baffles that divide the chambers **5** across which sound is attenuated.—JE

6,944,589

43.72.Ar VOICE ANALYZING AND SYNTHESIZING APPARATUS AND METHOD, AND PROGRAM

Yasuo Yoshioka and Jordi Bonada Sanjaume, assignors to Yamaha Corporation
13 September 2005 (Class 704/209); filed in Japan 9 March 2001

A routine is suggested which analyzes the voice sound into harmonic and inharmonic components, and which then extrapolates a harmonic spectrum envelope multiplied by a resonance curve (for the voice source) and a difference (error) spectrum envelope. It seems to be a sort of reinvention of linear prediction in that a parametric representation based on this procedure is subsequently computed and used for analysis and synthesis, but it is hard to be certain without exploring the mathematical equations presented in the patent more thoroughly.—SAF

6,950,531

43.66.Vt INDUSTRIAL HEARING PROTECTION AND COMMUNICATION ASSEMBLY

Tom Rickards, assignor to Energy Telecom, Incorporated
27 September 2005 (Class 381/381); filed 9 January 2003

An assembly is described consisting of safety glasses, a boom microphone for communication, and foam insert hearing protectors. The hearing protectors are attached with a flexible cable, that may be hollow, permitting sound to be transmitted from a receiver mounted on one of the temple pieces.—JE

6,947,539

43.72.Fx AUTOMATED CALL ROUTING

Jeffrey P. Graham, Franconia, New Hampshire *et al.*
20 September 2005 (Class 379/219); filed 14 August 2002

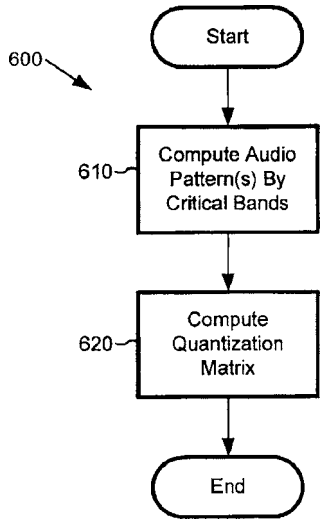
This patent describes a method for augmenting a voice-enabled call routing system with decision procedures based on prior information about the caller's requests. It is a useful application of a Bayesian philosophy that seeks to incorporate prior probabilities of a caller's behavior from outside the speech realm itself. The method simply uses this prior information to weight the probabilities of hypothesized utterances generated by the recognition component.—SAF

6,934,677

43.72.Gy QUANTIZATION MATRICES BASED ON CRITICAL BAND PATTERN INFORMATION FOR DIGITAL AUDIO WHEREIN QUANTIZATION BANDS DIFFER FROM CRITICAL BANDS

Wei-Ge Chen *et al.*, assignors to Microsoft Corporation
23 August 2005 (Class 704/200.1); filed 14 December 2001

To reduce overhead and required bitrate, quantization matrices are computed directly from critical band patterns using the same frequency coefficients in the auditory model that the encoder compresses, using



identical quantization matrices for sum and difference channels in stereo processing, and using a parametric compressor for low-bitrate applications.—DAP

6,947,886

43.72.Gy SCALABLE COMPRESSION OF AUDIO AND OTHER SIGNALS

Kenneth Rose *et al.*, assignors to The Regents of the University of California
20 September 2005 (Class 704/200.1); filed 21 February 2003

Quantization is performed for more than one layer to reduce the required bitrate for audio coding for a given sound quality with little computational cost. The result is that scalability of scalar quantizers is extended to systems with nonuniform base-layer quantization. A scalable, mean-square-error-based, companded quantizer may be used for both a base layer and one or more error reconstruction layers.—DAP

6,934,680

43.72.Ja METHOD FOR GENERATING A STATISTIC FOR PHONE LENGTHS AND METHOD FOR DETERMINING THE LENGTH OF INDIVIDUAL PHONES FOR SPEECH SYNTHESIS

Martin Holzapfel, assignor to Siemens Aktiengesellschaft
23 August 2005 (Class 704/245); filed in Germany 7 July 2000

Contrary to the title, this patent does not deal with the actual measurement of speech sound (phone) durations. That is to be done manually or perhaps using an HMM-based method. Measured phone durations are averaged in matching contexts of three and four phones in length, computing average durations and standard deviations. During synthesis, a candidate duration is based on the corresponding context length, if available, and, if not, computed from the statistics of similar contexts. The duration is then adjusted for the desired speaking rate.—DLR

6,952,676

43.72.Kb VOICE RECOGNITION PERIPHERAL DEVICE

William F. Sherman, Rancho Santa Margarita, California
4 October 2005 (Class 704/270); filed 11 July 2001

This patent claims priority over something that doesn't exist even in theory. The author describes the idea, completely free of encumbrances like detailed methods, of a ''peripheral device'' which would voice-enable a personal digital assistant. And that's it. I guess somebody ought to patent the Starship Enterprise next.—SAF

6,931,105

43.72.Ne CORRELATING CALL DATA AND SPEECH RECOGNITION INFORMATION IN A TELEPHONY APPLICATION

Christopher Ryan Groves and Kevin James Muterspaugh, assignors to International Business Machines Corporation
16 August 2005 (Class 379/88.03); filed 19 August 2002

The success rate for the recognition of numbers and possibly spelled-out words by a voice response system is arguably improved by checking wider sources of information when the present recognition confidence is low. Information about the present caller could include caller-ID info, ANI info, the called number, perhaps an Internet address, and historical values of any of those information sources from prior calls from the same caller.—DLR

6,931,263

43.72.Ne VOICE ACTIVATED TEXT STRINGS FOR ELECTRONIC DEVICES

Michael F. Boucher and Bart A. Smudde, assignors to Matsushita Mobile Communications Development Corporation of U.S.A.
16 August 2005 (Class 455/563); filed 4 March 2002

The system patented here appears to be a straightforward application of speech recognition technology used to control an electronic device. The input speech is converted to some form of unique string, either text or phonetics, which is, in turn, converted into a control command for the attached device. Such systems have been patented many times over. Here, the claims require that various embodiments include specific portions of the overall system, particularly involving certain variations in the processor software.—DLR

6,931,384

43.72.Ne SYSTEM AND METHOD PROVIDING UTILITY-BASED DECISION MAKING ABOUT CLARIFICATION DIALOG GIVEN COMMUNICATIVE UNCERTAINTY

Eric J. Horvitz and Timothy S. Paek, assignors to Microsoft Corporation
16 August 2005 (Class 706/45); filed 10 April 2001

In order to improve the reliability of human-computer interaction, the patented system would engage in a dialog with the user and would apply Bayesian analysis methods to determine the best course of action at every point. Multiple user-input streams would be considered, such that the system could respond to gestures, facial expressions, or other nonverbal inputs. Other lofty ideals are mentioned, such as cost/benefit analysis and inferring

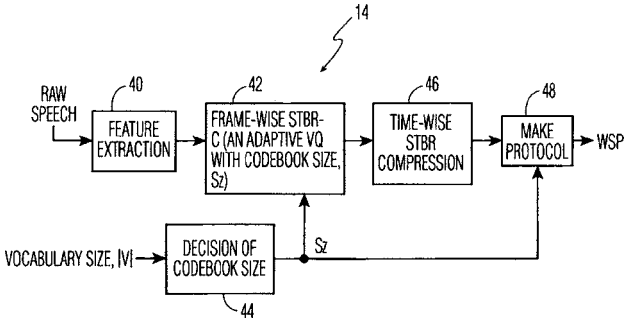
the user's goals. The brief patent does not describe software, training, or any other such down-to-Earth realities.—DLR

6,934,678

43.72.Ne DEVICE AND METHOD FOR CODING SPEECH TO BE RECOGNIZED (STBR) AT A NEAR END

Yin-Pin Yang, assignor to Koninklijke Philips Electronics N.V.
23 August 2005 (Class 704/221); filed 25 September 2000

To reduce required capacity for a distributed speech recognition system, the number of bits in the codebook size used to represent feature



vectors is dynamically adjusted, depending on the specific dialog or vocabulary size. The tradeoff between bitrate and expected recognition rate is optimized by minimizing a cost function.—DAP

6,934,683

43.72.Ne DISAMBIGUATION LANGUAGE MODEL

Yun-cheng Ju and Fileno A. Alleva, assignors to Microsoft Corporation
23 August 2005 (Class 704/257); filed 31 January 2001

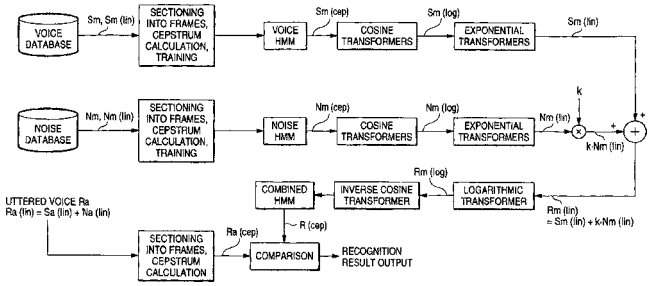
The discussion portion of this patent includes wide-ranging coverage of the value of good statistics in language modeling and some details of both *n*-gram and syntactic models, dealing primarily with the Chinese and Japanese languages. The claims cover the broad field of the use of language models for speech recognition of all Asian languages, including *n*-gram as well as syntactic models.—DLR

6,937,981

43.72.Ne VOICE RECOGNITION SYSTEM

Hiroshi Seo *et al.*, assignors to Pioneer Corporation
30 August 2005 (Class 704/246); filed in Japan 18 September 2000

To improve speech recognition in the presence of environmental noise and nonlinear distortion, a combined hidden Markov model is formed in the



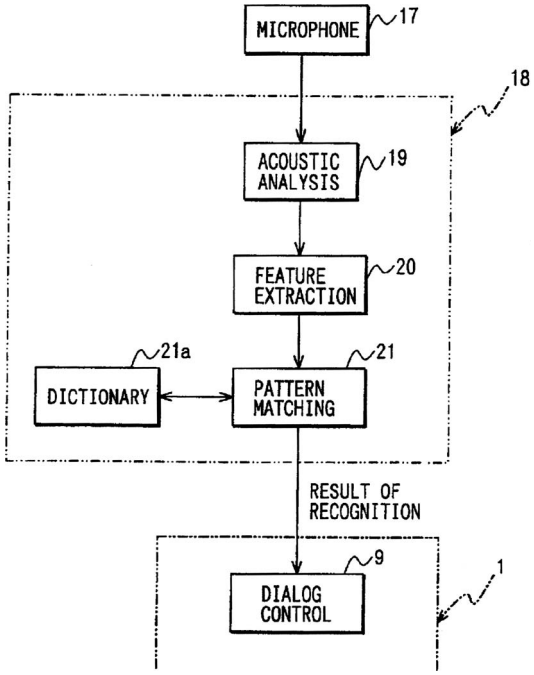
cepstral domain. Recognition is performed by comparing the spoken utterances with a noise model combined with an acoustic model.—DAP

6,937,982

43.72.Ne SPEECH RECOGNITION APPARATUS AND METHOD USING TWO OPPOSITE WORDS

Norihide Kitaoka and Hiroshi Ohno, assignors to Denso Corporation
30 August 2005 (Class 704/252); filed in Japan 21 July 2000

A technique is proposed for preventing an external device from performing an erroneous operation due to a misrecognition by a speech recognition system of an input as a wrong word in the stored vocabulary.



Reference patterns are established and are used with LPC and hidden Markov models for pattern matching to distinguish between similar sounding words in the vocabulary. Applications include car navigation systems with voice command input.—DAP

6,952,674

43.72.Gy SELECTING AN ACOUSTIC MODEL IN A SPEECH RECOGNITION SYSTEM

Richard A. Forand, assignors to Intel Corporation
4 October 2005 (Class 704/243); filed 7 January 2002

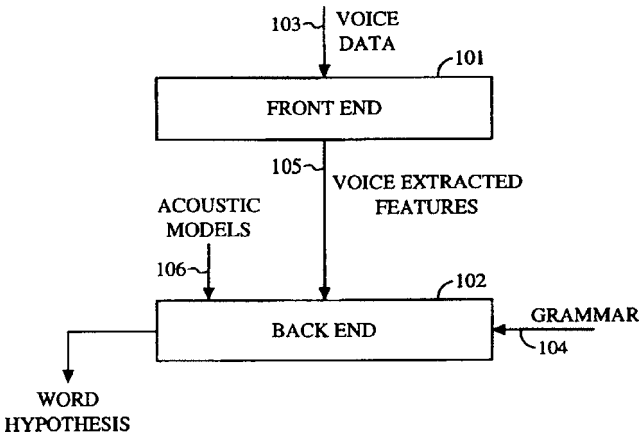
To improve the convenience of use and speed of adjustable, acoustic-model speech recognition systems, during setup, the acoustic model for the voice most closely representing the user's voice is selected from several stored sample voices.—DAP

6,941,265

43.72.Ne VOICE RECOGNITION SYSTEM METHOD AND APPARATUS

Ning Bi *et al.*, assignors to Qualcomm Incorporated
6 September 2005 (Class 704/246); filed 14 December 2001

A speech recognition system is provided with programmable front-end processing that can receive configuration files at different times. The stated advantage is that different back-end speech recognition systems can be



VOICE RECOGNITION SECTIONS

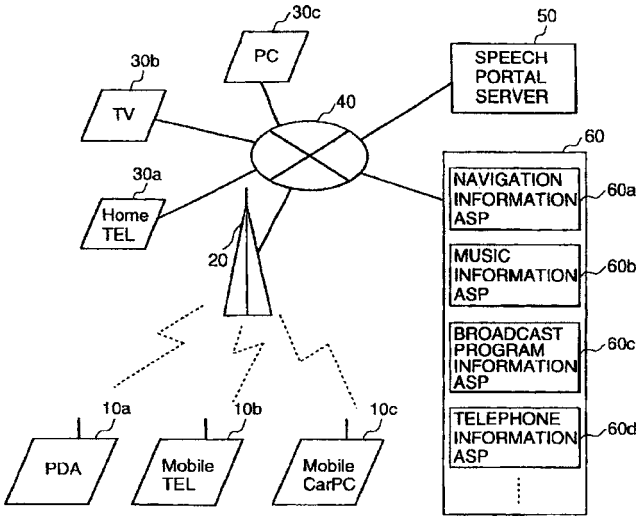
accommodated in which the extracted front-end features are compared with the grammar input to create a list of words and associated probabilities.—DAP

6,944,593

43.72.Ne SPEECH INPUT SYSTEM, SPEECH PORTAL SERVER, AND SPEECH INPUT TERMINAL

Soshiro Kuzunuki *et al.*, assignors to Hitachi, Limited
13 September 2005 (Class 704/270.1); filed in Japan 2 October 2001

A speech input/output system, accessed from a mobile or fixed terminal, provides maps, music, TV broadcast programming, and telephone



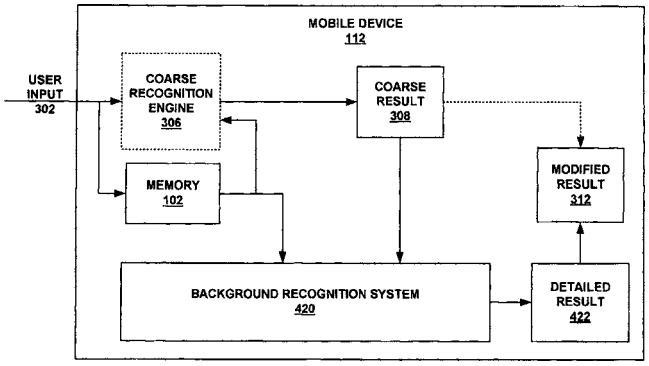
information. A speech recognizer checks inputs against a text dictionary and separates them into command text or object text.—DAP

6,950,795

43.72.Ne METHOD AND SYSTEM FOR A RECOGNITION SYSTEM HAVING A VERIFICATION RECOGNITION SYSTEM

Yoon Kean Wong, assignor to Palm, Incorporated
27 September 2005 (Class 704/231); filed 11 October 2001

To provide timely recall of information such as telephone numbers from mobile devices, or to access recorded information such as names,



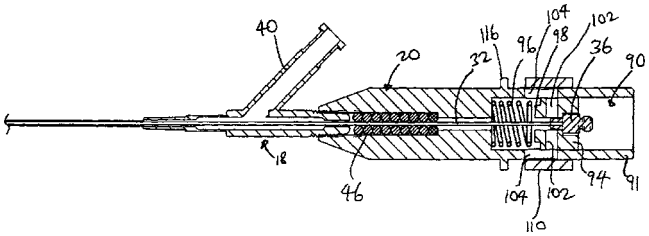
dates, and conversations, computer-implemented speech recognition is performed. A coarse recognition is first performed and results displayed, followed by a detailed recognition based on further user input.—DAP

6,942,620

43.80.Qf CONNECTOR FOR SECURING ULTRASOUND CATHETER TO TRANSDUCER

Henry Nita and Martinos Tran, assignors to Flowcardia Incorporated
13 September 2005 (Class 600/459); filed 19 September 2003

A connector is described for attaching a catheter to an ultrasonic transducer, which prevents frequency shifts and minimizes the mechanical



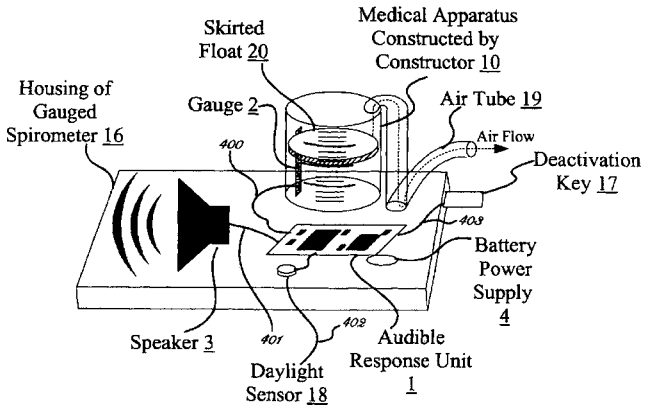
impact on the handling area during a medical procedure. The system features a transducer housing, an ultrasound transmission member, and a catheter knob with its proximal end coupled to the distal end of the housing.—DRR

6,942,625

43.80.Qf INCENTIVE SPIROMETRY DEVICES BY THE EMPLOYMENT OF VERBAL SIMULATED HUMANLIKE VOICES

Terry Keith Bryant, Singer Island, Florida
13 September 2005 (Class 600/538); filed 26 March 2004

A voice-prompting system is added to an incentive spirometer medical apparatus, a disposable plastic device used to rehabilitate lungs after surgery or similar type of event. The spirometer generally consists of a plastic bell jar containing a float that rises when air is inhaled through a tube. By inhaling the air, the patient attempts to reach different volumes that are measured by the float rising inside the jar. The patent covers an ancillary aid for providing voiced audible commands, encouraging phrases, responses, and guidance through the use of microchips, micro controllers, integrated



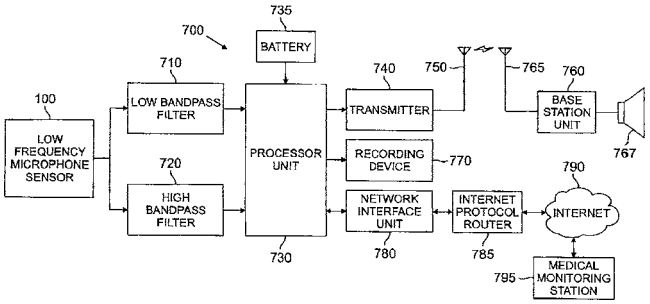
circuit controllers, and tiny loudspeakers, all of which are battery powered.—DRR

6,947,565

43.80.Qf PHYSIOLOGICAL CONDITION MONITORS UTILIZING VERY LOW FREQUENCY ACOUSTIC SIGNALS

Michael E. Halleck *et al.*, assignors to iLife Solutions, Incorporated
20 September 2005 (Class 381/67); filed 1 July 2002

The goal of this device is to assess the physiological condition of a patient, including cardiac and pulmonary activity, movement and position orientation of a body, and other types of physiological events entailing very-low-frequency signals in the range of 0.1–30 Hz. The device consists of a microphone positioned within a chamber. Mechanical energy from outside



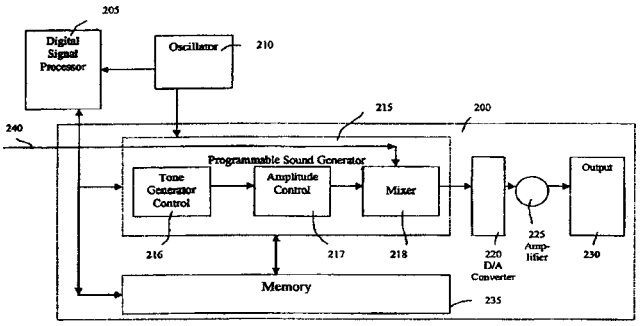
the chamber, such as pulmonary or cardiac activity, causes the chamber walls to move, resulting in low-frequency signals. The microphone is designed to detect extremely low-frequency waves in the fluid within the chamber. The signals obtained can be relayed to a monitoring station.—DRR

6,947,780

43.80.Qf AUDITORY ALARMS FOR PHYSIOLOGICAL DATA MONITORING

Tom D. Scharf, assignor to Dolphin Medical, Incorporated
20 September 2005 (Class 600/323); filed 31 March 2003

A system is described here for determining a patient's physiological status so that a health care provider can effectively monitor the patient by auditory alarms. For example, blood flow characteristics can be monitored without the necessity of maintaining visual contact. In one embodiment, at



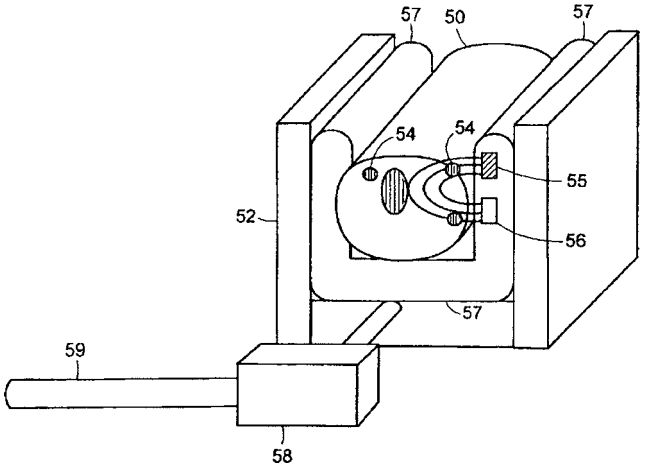
least one physiological characteristic is measured by a physiological data monitoring system consisting of a sensor unit and a monitoring unit. The monitoring unit contains a sound generator module to produce various signals indicative of physiological parameters and a memory module that stores commands for the programmable sound generator module.—DRR

6,947,781

43.80.Qf VIBRATORY VENOUS AND ARTERIAL OXIMETRY SENSOR

Haruhiko H. Asada and Phillip Shaltis, assignors to Massachusetts Institute of Technology
20 September 2005 (Class 600/335); filed 12 December 2003

This device is intended to apply an external perturbation to monitor a patient's blood constituents and to continuously measure the arterial and venous oxygen saturations. This involves the following steps. (1) Periodic vibration of a particular frequency is induced in the limb of a patient in such a manner as to excite a resonant response in a specified blood vessel. (2) The



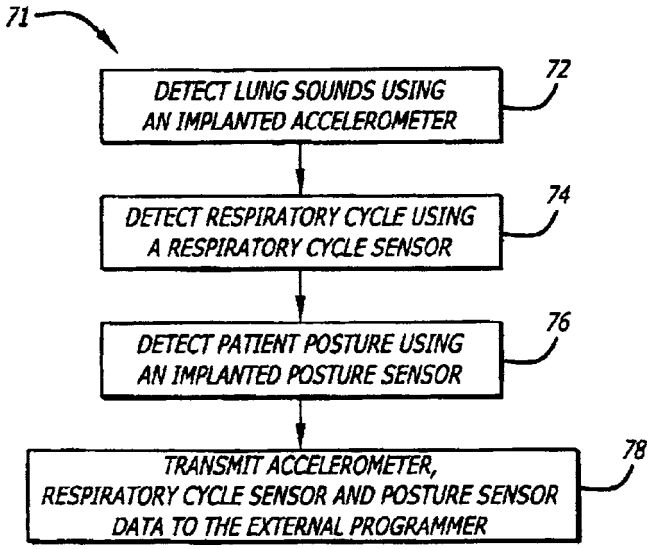
limb of the patient is illuminated with a light source. (3) Light is detected that has traversed the specified blood vessel and has generated a defining plethysmographic signal. (4) The synchronously detected plethysmographic signal discriminates a response attributable to the specified blood vessel.—DRR

6,949,075

43.80.Qf APPARATUS AND METHOD FOR DETECTING LUNG SOUNDS USING AN IMPLANTED DEVICE

John D. Hatlesad *et al.*, assignors to Cardiac Pacemakers, Incorporated
27 September 2005 (Class 600/586); filed 27 December 2002

A sensor is implanted adjacent to a pulmonary system to transmit signals resulting from lung sounds. A controller processes the signals to detect adventitious lung sounds. A respiratory cycle sensor operating in



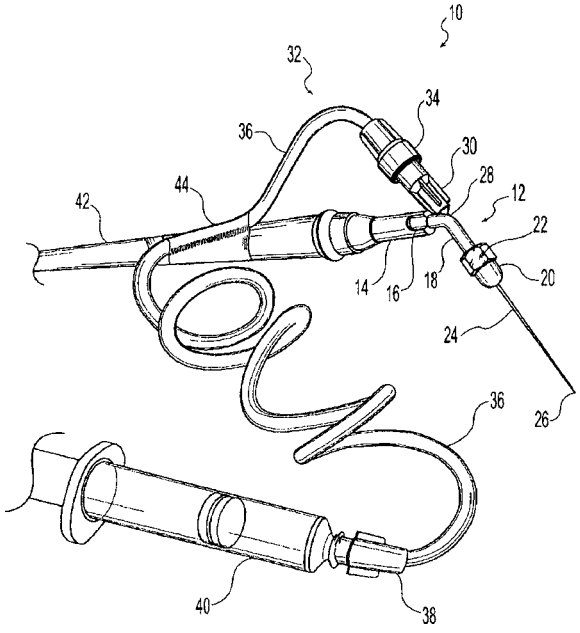
conjunction with the implanted lung sound sensor classifies an adventitious lung sound on the basis of its time of occurrence within the respiratory cycle. A posture sensor (which detects if a patient is lying down or sitting up) may provide additional information as to the severity of lung congestion.—DRR

6,948,935

43.80.Sh ULTRASONIC DENTAL DEVICE

John Nusstein, assignor to The Ohio State University
27 September 2005 (Class 433/119); filed 30 December 2002

A new twist in the procedure of root canal therapy is provided by this device for debriding the root canal space. The device incorporates a modified shaft assembly and a stainless steel hypodermic needle that concentrates ultrasound energy at its tip. The shaft assembly has a threaded housing for



attaching the shaft to an ultrasonic generator. A hub attached to the end of an angled shaft opposite the threaded housing reduces the diameter of the bore passing through its length. The needle is inserted into an aperture in the outer wall of the shaft and passes through the bore in the angled shaft to emerge through the hub. The dentist can adjust the working length of the needle.—DRR

6,945,938

43.80.Vj SYSTEMS AND METHODS FOR EVALUATING OBJECTS WITH AN ULTRASOUND IMAGE

Sorin Grunwald, assignor to Boston Scientific Limited
20 September 2005 (Class 600/443); filed 16 December 2002

An object is assumed to be defined by statistical properties that are consistently different from the corresponding statistical properties of the environment. If the statistical properties that are calculated at selected locations within the image fall within a range of values that represent the object, the locations are marked to indicate that they are positioned within the object. A border is then drawn around the object and the area calculated.—RCW

6,951,540

43.80.Vj ULTRASOUND IMAGING SYSTEM AND METHOD USING NON-LINEAR POST-BEAMFORMING FILTER

Emad S. Ebbini and Pornchai Phukpattaranont, assignors to Regents of the University of Minnesota
4 October 2005 (Class 600/437); filed 9 May 2003

A nonlinear post-beamforming filter based on a pth-order Volterra model is used to separate linear and harmonic components of ultrasound signals. Coefficients in the filter model are obtained by signal processing algorithms that are described. The filter-based approach extracts a broad band of frequencies for contrast enhancement while preserving image detail.—RCW

6,951,541

43.80.Vj MEDICAL IMAGING DEVICE WITH DIGITAL AUDIO CAPTURE CAPABILITY

Robert J. Desmarais, assignor to Koninklijke Philips Electronics, N.V.
4 October 2005 (Class 600/437); filed 3 December 2003

A sound channel associated with time-domain images is time-stamped using a system clock and recorded.—RCW

6,949,071

43.80.Vj METHOD FOR EXPLORING AND DISPLAYING TISSUE OF HUMAN OR ANIMAL ORIGIN FROM A HIGH FREQUENCY ULTRASOUND PROBE

Amena Saied *et al.*, assignors to Centre National de la Recherche Scientifique

27 September 2005 (Class 600/445); filed in France 12 January 1998

A computer-controlled, three-dimensional positioning system is used to position an ultrasound probe approximately orthogonal to the tissue to be scanned. Ultrasound pulse waves with a nominal frequency in the range of 30–100 MHz are emitted by the probe. Reflected signals are received and processed for display.—RCW

6,951,542

43.80.Vj METHOD AND APPARATUS FOR ULTRASOUND IMAGING OF A BIOPSY NEEDLE OR THE LIKE DURING AN ULTRASOUND IMAGING EXAMINATION

Barbara Greppi and Marino Cerofolini, assignors to ESAOTE S.p.A.

4 October 2005 (Class 600/443); filed 8 October 2002

Ultrasound images are formed using a set of transmit and receive parameters that are chosen to optimize imaging and visualization of a needle and to optimize imaging of the body into which the needle is inserted.—RCW

A review of finite-element methods for time-harmonic acoustics

Lonny L. Thompson[a)]
Department of Mechanical Engineering, Clemson University, Clemson, South Carolina 29634-0921

(Received 1 July 2005; revised 11 December 2005; accepted 12 December 2005)

State-of-the-art finite-element methods for time-harmonic acoustics governed by the Helmholtz equation are reviewed. Four major current challenges in the field are specifically addressed: the effective treatment of acoustic scattering in unbounded domains, including local and nonlocal absorbing boundary conditions, infinite elements, and absorbing layers; numerical dispersion errors that arise in the approximation of short unresolved waves, polluting resolved scales, and requiring a large computational effort; efficient algebraic equation solving methods for the resulting complex-symmetric (non-Hermitian) matrix systems including sparse iterative and domain decomposition methods; and a posteriori error estimates for the Helmholtz operator required for adaptive methods. Mesh resolution to control phase error and bound dispersion or pollution errors measured in global norms for large wave numbers in finite-element methods are described. Stabilized, multiscale, and other wave-based discretization methods developed to reduce this error are reviewed. A review of finite-element methods for acoustic inverse problems and shape optimization is also given. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2164987]

PACS number(s): 43.20.Bi, 43.20.Fn, 43.20.Rz [ADP] Pages: 1315–1330

I. INTRODUCTION

Finite-element methods (FEM) for time-harmonic acoustics governed by the reduced wave equation (Helmholtz equation) have been an active research area for nearly 40 years. Initial applications of finite-element methods for time-harmonic acoustics focused on interior problems with complex geometries including direct and modal coupling of structural acoustic systems for forced vibration analysis, frequency response of acoustic enclosures, and waveguides (Craggs, 1972; Gladwell, 1966; Nefske *et al.*, 1982; Petyt *et al.*, 1976; Young and Crocker, 1975; Zienkiewicz and Newton, 1969). In recent years, tremendous progress in the development of improved finite-element methods for time-harmonic acoustics including exterior problems in unbounded domains, which incorporate knowledge of wave behavior into the algorithm, combined with parallel sparse iterative and domain decomposition solvers, are moving the application of FEM into the higher frequency (wave number) regimes.

The exterior acoustics problem in unbounded domains presents a special challenge for finite-element methods. In order to use the FEM for exterior problems, the unbounded domain is usually truncated by an artificial boundary Γ yielding a bounded computational domain Ω; see Fig. 1. Reducing the size of the bounded domain reduces the computation cost, but must be balanced by the ability to minimize any spurious wave reflection with a computationally efficient and geometrically flexible truncation boundary treatment. The first complete finite-element approach for modeling acoustic radiation and scattering in unbounded domains appears in the impedance matching technique presented by Hunt *et al.* 1974, 1975. Recent numerical treatments including infinite elements, absorbing layers, local absorbing boundary conditions, and exact nonlocal boundary conditions have proven to be effective in handling acoustic scattering problems in unbounded domains, especially for large-scale problems requiring iterative and parallel solution methods and for modeling inhomogeneities and acoustic-structure interaction. The method of choice depends on the shape and complexity of the scattering object, inhomogeneities, frequency range, and resolution requirements, among other parameters.

A natural way of modeling the acoustic region exterior to a scattering/radiating object is to introduce a boundary element discretization of the surface S based on an integral representation of the exact solution in the exterior (Burton and Miller, 1971; Colton and Kreiss, 1983; Walsh *et al.*, 2004). Using the free-space Green's function (fundamental solution), the boundary element method (BEM) only requires surface discretization on S, reducing the d-dimensional problem to a $(d-1)$-dimensional one, and automatically satisfies the required Sommerfeld radiation condition at infinity (Sommerfeld, 1912). The BEM naturally incorporates surface impedance conditions but is limited in the ability to model complex, inhomogeneous regions. Application of the classical BEM for acoustic scattering requires solution of large, dense, complex linear systems due to the nonlocal support of the fundamental solution leading to high computational expense and storage requirements.

The finite-element method (FEM) is able to solve problems in nonhomogeneous media and allows for a natural coupling with complex structures. For exterior problems in unbounded domains, special techniques are required to reduce spurious reflection to a level below that of the discretization error. The numerical advantage of the FEM is that they lead to sparse matrices, which by avoiding calculations on zeros, significantly speed up computations and reduce memory requirements. Complexity estimates (Harari and Hughes, 1992) and numerical evidence (Burnett, 1994) have shown that domain-based methods such as the FEM are an

a)Electronic mail: lonny.thompson@ces.clemson.edu

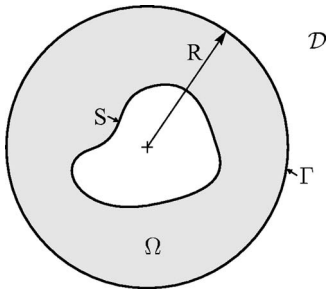


FIG. 1. Artificial truncation boundary Γ defining finite computational domain Ω for the exterior problem.

effective alternative to the BEM for exterior acoustics problems, especially for large systems due to the sparse structure of the resulting system matrices. With the recent developments in fast multipole methods which accelerate the calculation of matrix-vector products in iterative integral methods (Chen and Chen, 2004; Chew *et al.*, 1997, 2004; Darrigrand, 2002; Darve, 2000; Epton and Dembart, 1995; Fischer *et al.*, 2004; Greengard *et al.*, 1998; Gumerov and Duraiswami, 2004; Rockhlin, 1993; Sakuma and Yasuda, 2002; Schneider, 2003), the method with the best efficiency is less clear, yet the FEM retains the advantages of robustness and natural integration with other discrete models in coupled problems. It is also possible to couple the finite method with boundary integral methods (Fischer and Gaul, 2005), and other domain-based methods such as global Trefftz-based wave methods which are effective at high frequencies on moderate geometrical complexity (Van Hal *et al.*, 2003).

A difficulty of the standard Galerkin FEM applied to short-wave problems with wavelengths smaller than the geometrical parameters defining the domain has been the ability to accurately resolve oscillating wave solutions at higher frequencies (wave numbers). The difficulty of simultaneously achieving accuracy and efficiency at high wave numbers has been cited as one of the most challenging problems in scientific computation (Zienkiewicz, 2000). The failure to adequately control numerical dispersion errors not only inaccurately approximates the oscillatory part of the solution, but has a global pollution effect that builds up over the whole computational domain. The pollution effect is related to the loss of stability of the Helmholtz operator at large wave numbers. Due to challenges in accurately resolving short wave solutions at higher frequencies, many alternative and creative finite-element methods have been developed in the last decade including high-order methods, stabilized Galerkin methods, multiscale variational methods, and other wave-based discretization methods. A common theme of many of these improved finite-element methods is that they incorporate knowledge of wave behavior into the solution algorithm. Other recent improvements which are pushing the FEM into larger wave number regimes include the development of efficient iterative solvers with accelerating preconditioners, and domain decomposition parallel solution methods for the resulting sparse complex-symmetric (non-Hermitian) matrix systems arising from discretization of the Helmholtz operator.

In this paper, recent developments in finite-element methods for time-harmonic acoustics, including treatments of unbounded domains are reviewed. Topics include local and nonlocal Dirichlet-to-Neumann (DtN) nonreflecting boundary conditions, infinite elements, and absorbing layers for exterior problems; discretization methods which reduce numerical dispersion error arising in the approximation of short unresolved waves; efficient algebraic equation solving methods including sparse iterative and domain decomposition methods; a posteriori error estimates, adaptive methods; acoustic inverse problems, and shape optimization.

II. THE EXTERIOR PROBLEM IN UNBOUNDED DOMAINS

Let V be the domain of an object with boundary S. The exterior domain is defined by the unbounded region $\mathcal{R} = \mathbb{R}^3 \setminus V$. Time-harmonic acoustics is governed by the Helmholtz differential equation. For exterior problems defined on unbounded domains, solutions are required to satisfy the *Sommerfeld radiation condition* at infinity (Sommerfeld, 1912). The differential form of the boundary-value problem for exterior problems in unbounded domains may be stated for a general impedance surface condition as: Given wave number-dependent boundary data $g(\boldsymbol{x};k) \in \mathbb{C}$, $\beta(\boldsymbol{x};k) \in \mathbb{C}$; Find the complex-valued scalar field $u(\boldsymbol{x};k) \in \mathbb{C}$, such that

$$\nabla^2 u + k^2 u = 0, \quad \text{in } \mathcal{R} = \mathbb{R}^3 \setminus V \tag{1}$$

$$\frac{\partial u}{\partial n} + \beta u = g, \quad \text{on } S \tag{2}$$

$$\lim_{r\to\infty} r\left(\frac{\partial u}{\partial r} - iku\right) = 0. \tag{3}$$

Here, $u(\boldsymbol{x})$ represents the spatial part of the acoustic pressure or velocity potential, with wavenumber $k \in \mathbb{C}$, $\mathrm{Im}(k) \geqq 0$. The sign convention for the phase is $e^{-i\omega t}$, where $i = \sqrt{-1}$ and ω is the natural frequency. The normal derivative $(\partial u/\partial n) := \nabla u \cdot \boldsymbol{n}$ defines the gradient in the direction of the unit outward vector normal to S. In the above, $r = \|\boldsymbol{x}\|$ is a radius centered near the origin of the sound source. The *Sommerfeld radiation condition* (3) allows only outgoing waves proportional to $\exp(ikr)$ at infinity. The radiation condition requires that energy flux at infinity be positive, thus ensuring unique solutions.

Finite-element methods typically introduce an artificial boundary Γ, which divides the original unbounded domain into two regions: a bounded computational domain Ω discretized with the finite-element method and an infinite residual region $\mathcal{D} = \mathcal{R} \setminus \Omega$; see Fig. 1. Reducing the size of the bounded computational domain decreases the computational cost and memory storage. Methods for modeling the exterior complement $\mathcal{D} = \mathcal{R} \setminus \Omega$, i.e., the infinite region exterior to the artificial boundary Γ, can be divided into three main categories: local or nonlocal absorbing (nonreflecting) boundary conditions, infinite elements, and absorbing layers. Infinite element methods represent the exterior complement by assuming a radial approximation with outgoing wave behavior. Matched absorbing layers attempt to decay outgoing waves

in a relatively thin layer exterior to Γ. For the nonreflecting (absorbing) boundary conditions, the outgoing wave solution in $\mathcal{D}$ is represented by a relation of the unknown solution and its derivative on the artificial truncation boundary Γ. Options include matching exact analytical series solutions Hunt *et al.*, 1974, 1975, as used in the nonlocal Dirichlet-to-Neumann (DtN) map (Keller and Givoli, 1989), and various local approximations. The artificial boundary Γ is usually taken to be a surface defined in separable coordinates for efficiency, e.g., a sphere or spheroid. Formulations on nonseparable boundaries have also been developed; in this case the formulations are usually applied directly to the surface of the scatterer, thus completely avoiding discretization in Ω, e.g., Antoine, 2002; Antoine *et al.*, 1999; Shirron and Dey, 2002. Unbounded domain treatments may also be derived for acoustic waveguide problems (Bayliss *et al.*, 1983; Givoli, 1999; Murphy and Ching-Bing, 1989).

For absorbing boundary conditions, the originally unbounded exterior problem is replaced by an equivalent reduced problem defined on the bounded domain Ω: Find $u(\boldsymbol{x}) \in \mathbb{C}$, such that

$$\nabla^2 u + k^2 u = 0, \quad \text{in } \Omega \tag{4}$$

$$\frac{\partial u}{\partial n} + \beta u = g, \quad \text{on } S \tag{5}$$

$$\frac{\partial u}{\partial n} = \mathcal{B} u, \quad \text{on } \Gamma \tag{6}$$

where $\mathcal{B}$ is a linear operator called the *Dirichlet-to-Neumann* (DtN) map relating Dirichlet data to the outward normal derivative of the solution on Γ. The DtN operator $\mathcal{B}$ approximates the Sommerfeld radiation condition at a finite boundary Γ, and must satisfy the condition $\mathrm{Im}(u, \mathcal{B}u)_\Gamma \neq 0$ to ensure unique solutions. The DtN operator $\mathcal{B}$ is usually either a differential (local) or integral (nonlocal) operator, or combination of both. Physically, the DtN operator $\mathcal{B}$ represents radiation admittance relating pressure u (Dirichlet data) to normal velocity v_n which is proportional to the normal derivative $\partial u/\partial n = i\omega\rho v_n$ (Neumann data), on the truncation boundary Γ.

III. LOCAL ABSORBING BOUNDARY CONDITIONS

Absorbing boundary conditions should annihilate any spurious reflections at the artificial boundary (which are incoming). For local absorbing boundary conditions defined on a sphere, the development is based on the idea of annihilating radial terms in the Atkinson-Wilcox radial expansion in powers of $1/kr$ (Atkinson, 1949; Wilcox, 1956)

$$u(r, \theta, \varphi; k) = \frac{e^{ikr}}{ikr} \sum_{l=0}^{\infty} \frac{f_l(\theta, \varphi; k)}{(kr)^l}. \tag{7}$$

This expansion is valid for radius $r > r_0$, where r_0 is the radius of a spherical (or spheroidal) surface circumscribing the target/radiator, labeled S in Fig. 1, and any inhomogeneities of the domain Ω. Outside r_0, and in particular the radius R of the truncation surface, the exterior domain must be homogeneous and may not contain any objects/obstacles. Bayliss *et al.* (1982) showed that a sequence of local differential operators can be used to annihilate terms in this expansion. The first two local operators acting on the expansion for u, with their corresponding remainders are

$$G_1 u = \left(\frac{\partial}{\partial r} - \mathcal{B}_1 \right) u = O(1/kr)^3, \tag{8}$$

$$G_2 u = \left(\frac{\partial}{\partial r} + \frac{2}{r} - \mathcal{B}_1 \right) \left(\frac{\partial}{\partial r} - \mathcal{B}_1 \right) u = O(1/kr)^5, \tag{9}$$

where

$$\mathcal{B}_1 = ik - \frac{1}{r}. \tag{10}$$

In the case of the second-order operator G_2, the second-order radial derivative is replaced by second-order angular derivatives using the Helmholtz equation expressed in spherical coordinates. Setting the remainders to zero results in approximate local radiation boundary conditions which are easily implemented in standard finite-element methods. The corresponding local BGT boundary conditions are defined by relating the radial (normal) derivative to Dirichlet data in the form of the differential map,

$$\frac{\partial u}{\partial r} = \mathcal{B}_j u, \tag{11}$$

where for $j=1$, the first-order operator $\mathcal{B}_1$ is defined in (10) and for $j=2$, the second-order BGT operator is

$$\mathcal{B}_2 = \mathcal{B}_1 + \frac{1}{2\mathcal{B}_1} \Delta_\Gamma. \tag{12}$$

The second-order angular derivatives appearing in (12) are defined by

$$\Delta_\Gamma u := \nabla_\Gamma \cdot \nabla_\Gamma u = \frac{1}{r^2 \sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial u}{\partial \theta} \right) + \frac{1}{r^2 \sin^2\theta} \frac{\partial^2 u}{\partial \varphi^2},$$

where

$$\nabla_\Gamma := \frac{1}{r} \frac{\partial}{\partial \theta} \boldsymbol{e}_\theta + \frac{1}{r \sin\theta} \frac{\partial}{\partial \varphi} \boldsymbol{e}_\varphi.$$

The corresponding weak (variational) form of the boundary value problem with the $\mathcal{B}_2$ BGT operator is: Find the trial solution u, such that, for all test functions w

$$B(w,u) - B_\Gamma(w,u) = F(w), \tag{13}$$

where

$$B(w,u) := \int_\Omega (\nabla w \cdot \nabla u - k^2 wu) d\boldsymbol{x} + \int_S \beta wu \mathrm{d}\boldsymbol{s},$$

$$B_\Gamma(w,u) := \int_\Gamma \mathcal{B}_1 wu \mathrm{d}\Gamma - \int_\Gamma \frac{1}{2\mathcal{B}_1} \nabla_\Gamma w \cdot \nabla_\Gamma u \mathrm{d}\Gamma,$$

$$F(w) := \int_S wg \mathrm{d}\boldsymbol{s},$$

with differential surface area $\mathrm{d}\Gamma = R^2 \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\varphi$ on a spherical truncation boundary of radius R. The $\mathcal{B}_1$ and $\mathcal{B}_2$ operators both satisfy a required uniqueness condition, $\mathrm{Im}[B_\Gamma(u,u)] > 0$ (or <0), for all u evaluated on Γ, $u \neq 0$ (Grote and Keller, 1995; Harari and Hughes, 1992). The differential operators $\mathcal{B}_j$, $j=1,2$, are relatively simple to implement and retain the local sparse structure of the finite-element method. The local condition $\mathcal{B}_2$ is preferred since, for a fixed radius, it is more accurate compared to $\mathcal{B}_1$, as can be seen from the orders of the remainders in (9). Direct finite-element implementation of high-order operators $\mathcal{B}_j$, $j \geqslant 3$, are problematic in conventional finite-element methods since regularity in angular derivatives higher than standard $C^0(\Gamma)$ are required (Givoli *et al.*, 1997).

Conventional finite-element methods partition the computational domain Ω into nonoverlapping subdomains (elements) Ω_e with continuous piecewise polynomials. In the standard h-version, basis (shape) functions $N_i(\boldsymbol{x})$, associated with element nodes are C^0 continuous interpolation functions with compact support. The continuous approximation is written as the linear combination,

$$u^h(\boldsymbol{x}) = \sum_{i=1}^{N_{\mathrm{dof}}} N_i(\boldsymbol{x}) d_i = \boldsymbol{N}^T(\boldsymbol{x}) \boldsymbol{d}, \tag{14}$$

where $\boldsymbol{N} \in \mathbb{R}^{N_{\mathrm{dof}}}$ is a column vector of standard C^0 basis functions, and $\boldsymbol{d} \in \mathbb{C}^{N_{\mathrm{dof}}}$ is a column vector containing the N_{dof} unknown nodal values $d_i = u^h(\boldsymbol{x}_i)$, where $u^h(\boldsymbol{x}_i)$ is the approximation of the solution u at node $\boldsymbol{x}_i$. Using a standard Galerkin finite-element approximation, test (weighting) functions w^h are expressed as a linear span of the same basis functions. Substitution into (13) leads to the sparse, complex-symmetric (non-Hermitian) linear algebraic system,

$$\boldsymbol{Z}\boldsymbol{d} = \boldsymbol{f}, \quad \boldsymbol{Z} = (\boldsymbol{S} - k^2 \boldsymbol{M} - \boldsymbol{K}_\Gamma), \tag{15}$$

with matrix coefficients

$$(\boldsymbol{S})_{ij} = \int_\Omega \nabla N_i \cdot \nabla N_j \mathrm{d}\boldsymbol{x} + \int_S \beta N_i N_j \mathrm{d}\boldsymbol{s},$$

$$(\boldsymbol{M})_{ij} = \int_\Omega N_i N_j \mathrm{d}\boldsymbol{x},$$

$$(\boldsymbol{K}_\Gamma)_{ij} = \int_\Gamma \mathcal{B}_1 N_i N_j \mathrm{d}\Gamma - \int_\Gamma \frac{1}{2\mathcal{B}_1} \nabla_\Gamma N_i \cdot \nabla_\Gamma N_j \mathrm{d}\Gamma.$$

The excitation vector

$$(\boldsymbol{f})_i = \int_S N_i g \mathrm{d}\boldsymbol{s}, \tag{16}$$

is in general complex-valued, and wave number dependent. The contribution from the local absorbing conditions are found in the k-dependent sparse matrix $\boldsymbol{K}_\Gamma$ associated with node points on the boundary Γ. Further details of the finite-element implementation such as element mapping, integration, and assembly of element arrays are found in standard finite-element textbooks, e.g., Hughes, 2000; Petyt, 1990. The local support of the element shape functions gives the advantage of being able to handle complex geometries and producing sparse matrices which are solved efficiently by avoiding storage and computation of zero coefficients. For large three-dimensional problems at high wave numbers k, accurate resolution requires a large number of elements leading to large sparse matrices. In this case, iterative solvers are preferred over direct factorization methods due to the lower memory requirements and parallel computing performance.

Other local absorbing conditions which attempt to annihilate incoming waves include the Enquist and Majda (1977) and Feng (1983) conditions. In a numerical study by Shirron (1998), the accuracy of the BGT conditions are the most accurate, especially for low modes and tight boundaries. The first- and second-order BGT conditions have been widely used Bossut and Decarpigny, 1989; Kechroud *et al.*, 2004; Tezaur *et al.*, 2001 and have been generalized to spheroidal (Grote and Keller, 1995) and arbitrary convex surfaces (Antoine *et al.*, 1999). The use of spheroidal or rectangular coordinates allows the artificial boundary to obtain a tight fit around elongated objects. For spheroidal, and other convex shapes, the conditions tend to lose accuracy for higher wave numbers (Tezaur *et al.*, 2002). Low-order approximate conditions require careful placement when computing the response over a range of frequencies; the size of the computational domain and the mesh density must be carefully selected to achieve a prescribed accuracy. If not placed sufficiently far from the radiating/scattering object, low-order local approximate absorbing boundary conditions may produce large spurious reflections which can pollute the entire numerical solution.

Complexity estimates show that it is usually more efficient to use high-order accurate absorbing conditions which enable smaller computational domains. The development of high-order local boundary conditions for which the order can be easily increased to a desired level are usually based on using auxiliary variables to eliminate higher-order derivatives (Givoli, 2004; Hagstrom, 1999; Hagstrom and Hariharan, 1998; Huan and Thompson, 2000; Thompson *et al.*, 2001; van Joolen *et al.*, 2003). While generally derived for the time-dependent case, time-harmonic counterparts are readily implemented with time derivatives replaced by $i\omega$, $\omega = kc$, where c is wave speed.

IV. THE DTN NONREFLECTING BOUNDARY CONDITION

An alternative to high-order local absorbing conditions are nonlocal DtN nonreflecting boundary conditions. The conceptual foundation and experimental validation for the DtN finite-element method are presented for both acoustic radiation and scattering in the impedance matching technique of Hunt *et al.*, 1974, 1975. In Hunt (1974), the relationship between the pressure (Dirichlet data) and its normal derivative (Neumann data) on a spherical surface Γ is obtained using the surface Helmholtz integral equation with boundary condition $\partial u/\partial n = i\omega\rho v_n$, and then by expanding the analytical solution for the pressure u and normal velocity v_n in terms of spherical harmonics and matching coefficients. In Keller and Givoli (1989) the DtN map on a sphere of radius R is constructed directly and implemented in the standard Galerkin finite-element method by expanding the outgoing acoustic field in a spherical harmonic series,

$$u(r,\theta,\varphi) = \sum_{n=0}^{N-1} \frac{h_n(kr)}{h_n(kR)} \sum_{m=-n}^{n} u_{nm} Y_{nm}(\theta,\varphi), \tag{17}$$

with coefficients

$$u_{nm} = (u, Y_{nm})_{\mathcal{S}} := \int_0^{2\pi} \int_0^{\pi} u(R,\theta,\varphi) Y_{nm}^* d\mathcal{S}. \tag{18}$$

In the above, $d\mathcal{S} = \sin\theta\, d\theta\, d\varphi$ is the differential surface element on the unit sphere $\mathcal{S}$, parametrized by $0 < \theta < \pi$, $0 < \varphi < 2\pi$, and

$$Y_{nm}(\theta,\varphi) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\varphi}$$

are angular spherical harmonics such that $Y_{n,(-m)} = (-1)^m Y_{nm}^*$; the star denotes complex conjugate; and $h_n(kr)$ are outgoing radial spherical Hankel functions. The DtN map is then obtained by evaluating the normal derivative of (17) on the boundary at $r=R$, and implemented as a "natural" boundary condition with standard finite-element basis functions for u on the surface Γ. Givoli (1999) recognized that the DtN finite-element method could be generalized to other boundary value problems with infinite domains. The operator is nonlocal since the coefficients u_{nm} in (18) require integration over the whole surface. The DtN map exactly represents all harmonics in the solution up to the number of terms included in the truncated series expansion as measured by N. For higher harmonics $n > N-1$, the truncated DtN models the boundary Γ with the homogeneous Neumann condition $\partial u/\partial r = 0$ at $r=R$. As a consequence, nonunique solutions may result when k^2 matches an interior resonance associated with the Laplacian operator. Harari and Hughes (1992) showed that this difficulty can be eliminated by using a sufficient number of harmonics N. However, the restriction may require more terms in the DtN map than may be necessary to achieve a desired accuracy, leading to a potential for excessive computation.

This problem is circumvented if a modified truncated DtN operator (Grote and Keller, 1995) is used, $\mathcal{M}^* = (\mathcal{M}_N - \mathcal{B}_N) + \mathcal{B}$, where $\mathcal{B}$ is any computationally efficient approximation to the DtN operator with the uniqueness property $\mathrm{Im}(u, \mathcal{B}u)_\Gamma \neq 0$, and $(\mathcal{M}_N - \mathcal{B}_N)$ is the truncation of $\mathcal{M} - \mathcal{B}$ to the first N modes. The modified DtN condition provides unique solutions at all wave numbers irrespective of the number of harmonics N included in the series. Suitable operators $\mathcal{B}$ include local absorbing boundary conditions. The boundary condition $\mathcal{B}_2$ is preferred over $\mathcal{B}_1$, since it provides an improved matrix preconditioner for iterative solvers and gives more accurate solutions when the number of harmonics N used in the truncated DtN series is not sufficient to capture important modes $n > N-1$ in the solution. Applying the local $\mathcal{B}_2$ operator to (17) gives the modified DtN

$$\frac{\partial u}{\partial r} = \mathcal{B}_2 u + \sum_{n=2}^{N-1} \beta_n \sum_{m=-n}^{n} u_{nm} Y_{nm}(\theta,\varphi), \tag{19}$$

where

$$\beta_n = k\frac{h_n'(kR)}{h_n(kR)} + \frac{n(n+1)}{2\mathcal{B}_1 r^2} - \mathcal{B}_1. \tag{20}$$

The $\mathcal{B}_1$ modified DtN is a special case obtained by omitting the second term with $n(n+1)$ and starting the summation at $n=1$. Thompson and Pinsky (1996) recognized that these conditions also annihilate up to the first $N=2$ spherical modes corresponding to $n=0$ and $n=1$ in the expansion (17) and thus are equivalent to the first two localized DtN conditions derived in Harari and Hughes (1992).

Nonlocal conditions are very accurate, yet couple all solution unknowns on Γ, thus potentially rendering a full dense matrix with associated solution cost and memory requirements. However, if separable boundaries are utilized such as spheres or spheroids, a special structure in the resulting data structures may be exploited to avoid storage of a full dense matrix. Bayliss *et al.* (1985) appear to be the first to recognize that a global DtN radiation boundary condition formed by a harmonic expansion, and relating Dirichlet-to-Neumann data on a separable boundary can be split as a vector outer product which then can be used to perform matrix-by-vector multiplies in iterative solvers without the need for assembling a dense matrix.

The contribution of the DtN operator to the complex-symmetric (non-Hermitian) system matrix is defined by the admittance matrix,

$$\boldsymbol{K}_{\mathrm{dtn}} = R^2 \sum_{n=2}^{N-1} \beta_n \sum_{m=-n}^{n} \boldsymbol{c}_{nm} \boldsymbol{c}_{nm}^T, \tag{21}$$

where $\boldsymbol{c}_{nm} = (\boldsymbol{N}, Y_{nm})_{\mathcal{S}}$ are vectors of size equal to the number of unknowns on the truncation boundary Γ.

Due to the special structure of the DtN map defined on a separable boundary as a summation of vector outer products, we recognize that the Sherman-Morrison algorithm in conjunction with direct solvers may be used to preserve the sparsity of the finite-element equations with a series of rank-1 vector updates. For Krylov subspace iterative solvers the computationally intensive kernel is the repeated operation of matrix-by-vector products with vector iterates. The special structure of the DtN matrix $\boldsymbol{K}_{\mathrm{dtn}}$, as a summation of rank-1 vector updates can be exploited to avoid direct evaluation of

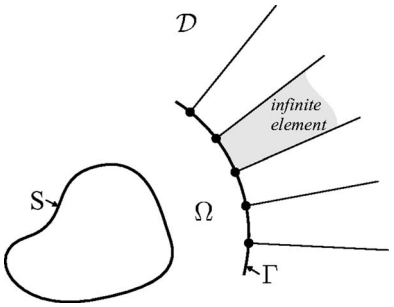


FIG. 2. Infinite element topology.

matrix-vector products with $\boldsymbol{K}_{\text{dtn}}$ with significantly reduced storage and cost (Oberai *et al.*, 1998). The matrix-vector product of the DtN operator can also be carried out at the element level, preserving standard element-based data structures (Malhotra and Pinsky, 1996). Oberai *et al.* 1998, Thompson *et al.* (1999) have shown that the local operator $\boldsymbol{K}_{\Gamma}$ provides a good approximation to the spectral properties of the complete system matrix which includes the DtN matrix $\boldsymbol{K}_{\text{dtn}}$, and thus can be used as an efficient preconditioner to accelerate convergence. Parallel iterative methods provide a means for dividing the problem into subsystems which when solved in parallel, provide compute time speedup, and for distributed-memory computer systems, the ability to scale up to very large systems. Ianculescu and Thompson (2003) showed that the symmetric outer-product structure of the DtN matrix $\boldsymbol{K}_{\text{dtn}}$ can be exploited to compute in parallel with one collective communication per iteration with a vector size equal to the number of harmonics included in DtN series expansion; the effect on the overall communication is roughly that of a relatively small dot-product interprocessor communication. Numerical studies reported in Ianculescu and Thompson (in press) show that due to the special structure, and ability to tightly fit around scattering objects with minimal spurious reflection, the nonlocal DtN condition can be implemented with significant overall cost savings compared to the local operators $\mathcal{B}_1$ and $\mathcal{B}_2$.

The extension of the $\mathcal{B}_2$ modified DtN map in spheroidal coordinates suitable for finite-element implementation is given in Thompson *et al.*, 1999. Grote and Keller (1995) derived a related modified DtN condition for spheres and spheroids using the second-order BGT operator in native form involving second-order radial derivatives (instead of angular derivatives), a form which is not suitable for standard finite-element approximation. Further details and discussions of the properties of DtN nonreflecting boundary conditions are given in Givoli (1999) and Thompson and Pinsky (2004).

V. INFINITE ELEMENTS

Infinite elements replace the nonreflecting boundary condition on Γ with a single layer of elements with infinite extent. The infinite elements are constructed with radial wave functions which automatically satisfy the Sommerfeld condition (3) at infinity; see Fig. 2. Test and trial solutions in the region Ω_X exterior to Γ are usually separated into radial and angular functions; for a spherical boundary Γ with radius R (Astley, 2000),

$$w = \sum_{\mu} \sum_{\nu} c_{\mu\nu} W_{\nu}(r) N_{\mu}(\theta, \varphi), \tag{22}$$

$$u = \sum_{\mu} \sum_{\nu} d_{\mu\nu} U_{\nu}(r) N_{\mu}(\theta, \varphi). \tag{23}$$

In the above, N_{μ} are angular basis functions which match the interior finite-element discretization on the surface Γ. Radial functions are defined to match the outgoing wave character of the radial expansion (7),

$$U_{\nu}(r) = R_{\nu}(\xi) e^{ik(r-R)}.$$

Here, the radial basis functions R_{ν} are polynomial functions in the inverse radius variable $\xi = (R/r)$, where $r := \|\boldsymbol{x}\| > R$ is the radial position exterior to Γ. Different definitions of the radial polynomial functions lead to changes in the conditioning of the resulting system matrix. Several alternatives for the radial test (weighting) functions have been proposed; the three most common choices are

$$W_{\nu}(r) = \begin{cases} U_{\nu}(r), & \text{Bettess – Burnett unconjugated} \\ U_{\nu}^{*}(r), & \text{Burnett conjugated} \\ \xi^2 U_{\nu}^{*}(r), & \text{Astley – Leis conjugated.} \end{cases}$$

Bettess (1977, 1992) pioneered the infinite element concept and selected the test function to be the same as the trial solution basis. Burnett (1994), Burnett and Holford 1998a, b extended the formulation to spheroidal and ellipsoidal coordinates, and was the first work to express the shape functions as separable tensor products of radial and transverse functions, resulting in improved performance and efficiency. A quantitative error analysis of the unconjugated infinite elements is given in Burnett and Holford (1998a). Alternatively, the weighting (test) function is conjugated. In the Astley-Leis infinite element (Astley *et al.*, 1998a) the conjugated weighting function is scaled by a geometric factor. It was later recognized that this formulation fits within the variational framework of Leis (1986). The unconjugated infinite element leads to matrix coefficients involving one-dimensional radial infinite integrals which are oscillatory and well-defined, and can be evaluated using high-order Gauss-Legendre quadrature. For the conjugated elements, the oscillatory plane-wave components cancel so that the radial coefficients may be integrated analytically in closed form, resulting in wave number-independent matrices which are proportional to ik and k^2, a feature which allows for a direct local time-dependent counterpart (Astley *et al.*, 1998b).

The unconjugated Burnett formulation, regardless of the definition for the radial function R_{ν}, gives the highest accuracy in the near field, yet exhibits instability and ill-conditioning for higher radial orders (Ihlenburg, 2000; Shirron and Babuska, 1998). For the Astley-Leis conjugated schemes, although less accurate in the near field, in the case of a spherical boundary, R_{ν} can be constructed so that the formulation remains stable and convergent in the far field (Astley, 2000; Astley and Hamilton, 2000, Dreyer and von

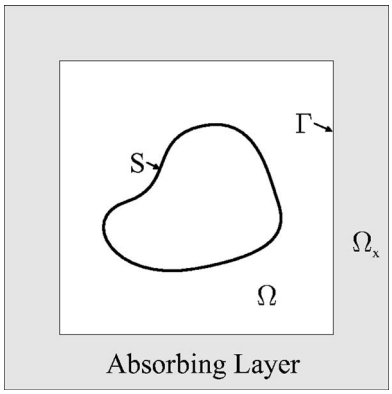


FIG. 3. Absorbing layer topology.

Estorff, 2003; Gerdes, 1998; Shirron and Babuska, 1998). However, the performance of both conjugated and unconjugated formulations deteriorates at larger wavenumbers and highly elongated artificial interfaces (Astley and Coyette, 2001).

VI. ABSORBING BOUNDARY LAYERS

The perfectly matched layer (PML) concept originally introduced by Berenger (1994) for electromagnetic waves is another option for modeling the far field for the exterior acoustics problem. The interface and PML are usually formulated in rectilinear Cartesian coordinates, allowing a tight fit around elongated objects (see Fig. 3), but can also be formulated in spherical and other general curvilinear coordinates. The idea is to introduce an exterior layer of finite thickness at an artificial interface such that outgoing plane waves are absorbed prior to reaching the outer-layer truncation boundary. By splitting the scalar field into nonphysical components satisfying equations which describe decaying waves, and proper selection of PML coefficients, plane-wave reflection for an arbitrary angle of incidence is theoretically eliminated.

In the absorbing layer Ω_X, splitting the field in Cartesian coordinates $\boldsymbol{x}=(x_1,x_2,x_3)=(x,y,z)$, leads to a modified Helmholtz equation with complex-valued anisotropic material properties suitable for standard finite-element implementation (Harari *et al.*, 2000; Turkel and Yefet, 1998; Wu *et al.*, 1997),

$$\nabla\cdot(\boldsymbol{D}\,\nabla u)+k^2 s u=0,\quad \text{in } \Omega_X. \tag{24}$$

The corresponding weak form in the complete computational region $\Omega\cup\Omega_X$ is

$$\int_{\Omega\cup\Omega_X}(\nabla w\cdot\boldsymbol{D}\,\nabla u-k^2 s w\cdot u)d\boldsymbol{x}=F(w). \tag{25}$$

In the above, $s=s_1s_2s_3$, and $\boldsymbol{D}=\mathrm{diag}\{s_2s_3/s_1,s_1s_3/s_2,s_1s_2/s_3\}$ is a diagonal, complex-valued material matrix, with coefficients $s_i(x_i)=1+(i\sigma_i)/k$, defined by a distribution of absorption functions $\sigma_i(x_i)$, $i=1,2,3$, usually taken to vary quadratically from a value of zero at the interface of the physical domain to a maximal value at the truncation of the layer. In the physical domain Ω, $s_i=1$, $i=1,2,3$. In layers normal to the x_1 direction, $\sigma_2=\sigma_3=0$. Only in corner regions are all σ_i values nonzero. Optimal placement of the interface, layer thickness, number of elements, and variation of absorption functions and their maximal value, which reduce error due to spurious reflection of decayed waves off the layer truncation boundary to be less than the discretization error, are open questions (Collino and Monk, 1998). The PML layer converges to perfect wave absorption as the thickness of the layer is increased (Tsynkov and Turkel, 2001). However, a compromise between a thin layer which requires a rapid variation of the absorption parameters and a thick layer which requires more elements is required (Turkel and Yefet, 1998).

VII. DISCRETIZATION METHODS FOR THE HELMHOLTZ EQUATION

A. Galerkin finite-element methods

Accuracy of finite-element approximations based on Galerkin's method is characterized by phase (dispersion) errors. In order to control dispersion, the element size must be adapted to the wave number such that the number of elements per wavelength is held below a resolution limit. The resolution is determined by the nondimensional wave number $kh=2\pi h/\lambda$, where λ is the wavelength and h is a measure of the element size. A discrete dispersion analysis can be used to quantify the limiting value on mesh resolution.

For a uniform mesh of finite elements with piecewise linear interpolation in one dimension, and neglecting boundary conditions, the sparse system matrix $\boldsymbol{Z}$ is tridiagonal; neglecting sources, each interior equation corresponds to a repeated finite difference stencil centered at a typical node point x_j of the form

$$Z_2u^h(x_{j-1})+2Z_1u^h(x_j)+Z_2u^h(x_{j+1})=0, \tag{26}$$

where $Z_1(kh)=1-(kh)^2/3$ and $Z_2(kh)=-(kh)^2/6-1$ are coefficients obtained by assembly of element matrices. The solution of the difference stencil admits homogeneous plane-wave solutions of the form $u^h(x_j)=u_0e^{i\tilde{k}x_j}$ where $\tilde{k}$ is an unknown numerical wave number. Substituting this solution into the difference stencil leads to a dispersion relation relating the numerical wave number $\tilde{k}$ to the continuous wave number k of the form, $\cos(\tilde{k}h)=-Z_1(kh)/Z_2(kh)$. For small kh, a Taylor series expansion reveals the dispersion error of order (Thompson and Pinsky, 1994)

$$(\tilde{k}-k)/k=-\frac{1}{24}(kh)^2+O(kh)^4. \tag{27}$$

The numerical wave number remains real valued corresponding to propagating waves, provided $|Z_1(kh)/Z_2(kh)|<1$, which requires the continuous wave number k to be bounded by the cutoff value, $kh\leqslant\sqrt{12}$, corresponding to a minimum resolution of $\lambda/h>2$, i.e., just under 2 elements per wavelength. Beyond this value, $\tilde{k}$ is complex valued, resulting in rapid amplitude decay (evanescent behavior).

B. Mesh resolution rules for low-order elements

Ihlenburg (1998), Ihlenburg and Babuska (1995) showed that the error measured in the usual L_2 integral norm, which characterizes averaged amplitude differences, is controlled by a wave number-dependent stability constant times the approximation error. For linear finite elements of size h,

$$\left(\int_\Omega |u^h - u|^2 dx\right)^{1/2} \leqslant (C_1 + C_2 kL)(kh)^2, \tag{28}$$

where L is a characteristic length scale, and C_1, C_2 are constants independent of the wave number and element size. The meaning of (28) is that the integral norm is controlled by a sum of two errors, where the first term $C_1 \cdot (kh)^2$ is proportional to the usual approximation error, and the second term $C_2 \cdot (kL)(kh)^2$ has an additional dependence on kL. The second term has sometimes been referred to as a "pollution" error related to a loss of stability at large wave numbers (Bayliss *et al.*, 1985). The second term increases even though the nondimensional wave number kh is held fixed with a constant number of elements per wavelength (λ/h =constant). In general, however, if the number of elements per wavelength λ/h is increased (kh decreased), it follows that the second term also decreases, thus reducing the error in amplitudes. To control local approximation error it is often suggested that one take at least ten ($\lambda/h > 10$) linear elements per wavelength for a phase error of a few percent, i.e., keep the element size below $h < \lambda/10$. However, to control amplitude error, a characteristic length scale of the domain, L, should be accounted for such that $(kL)(kh)^2 < P$, where P is an admissible pollution error determined from computational experience (Ihlenburg, 2003).

A simple way to improve both dispersion and pollution error for low-order elements is to use slightly underintegrated stiffness and mass matrices with special numerical quadrature rules. For 4-node bilinear quadrilateral and 8-node trilinear "brick" elements, use of quadrature evaluation points $\pm\sqrt{2/3}$ in each direction in the parent element with unit weights gives a higher-order accurate dispersion error from $O(kh)^2$ to $O(kh)^4$, (Challa, 1998). This same observation for 4-node quadrilateral elements was later rediscovered in Guddati and Yue (2004), where it was demonstrated numerically that the high-order accuracy is maintained for unstructured meshes. In Thompson and Kunthong (2005) it is shown that for 3-node linear triangle elements, the dispersion error is reduced from $O(kh)^2$ to $O(kh)^4$ by using three-quadrature points for the mass matrix $\boldsymbol{m}^e$ sampled at $(\xi, \eta, 1-\xi-\eta) = (1/3 + \sqrt{10}/6, 1/3 - \sqrt{10}/12, 1/3 - \sqrt{10}/12)$ in natural coordinates with three symmetric permutations and conventional weights $\Sigma_{j=1}^3 W_j = 1$, $W_j = 1/3$. For 4-node tetrahedral elements for three-dimensional analysis, using the special quadrature rule $(\xi, \eta, \zeta, 1-\xi-\eta-\zeta) = (1/4 + \sqrt{21}/8, 1/4 - \sqrt{21}/24, 1/4 - \sqrt{21}/24, 1/4 - \sqrt{21}/24)$, with multiplicity 4, and standard weights $\Sigma_{j=1}^4 W_j = 1$, $W_j = 1/4$, yields increased phase accuracy compared to standard Galerkin.

C. High-order approximation

Dispersion error can be minimized by using higher-order polynomial approximations, e.g., hp-version of FEM and spectral elements. The number of elements per wavelength to obtain a given discretization error depends strongly on the order of the element basis functions. A dispersion analysis similar to that outlined above for linear elements can be carried out for high-order polynomials of order $p \geqslant 2$, and after condensation of internal solution unknowns, a dispersion relation in the same form as the linear $p=1$ case is obtained. Thompson and Pinsky (1994) and Ihlenburg (1998) show that the relative phase error is $O(kh)^{2p}$. The cutoff value prior to evanescent behavior grows with the increase of approximation order p; however, before reaching this value the numerical wave number is complex valued on small "stopping band" intervals (Thompson and Pinsky, 1995). As a result, kh should be kept below the first cutoff value of $kh \leqslant \sqrt{12}$. Ihlenburg (1998) showed that the error measured in the $H^1(\Omega)$ global norm is order $kL(kh/2p)^{2p}$, and thus for quadratic and higher-order elements $p \geqslant 2$, and resolved waves such that kh/p is small, the dispersion is relatively small. For high wave numbers $kh \gg 1$, Ainsworth (2004) has shown that dispersion error is virtually eliminated when p is increased to a regime in which the error decays at a super-exponential rate, such that $2p+1 > kh + c(kh)^{1/3}$, where c is a user-defined constant; $c=2$ is suggested. Since higher-order elements generally provide greater computational efficiency, fewer degrees of freedom are generally needed to achieve a given discretization error for oscillatory wave solutions (Deraemaeker *et al.*, 1999; Dey, 2003; Thompson and Pinsky, 1994). For high-order quadrilateral elements with quadratic Lagrange polynomial basis functions of order $p=2$, high-order accuracy is achieved by evaluating the stiffness and mass matrices with special quadrature points $\xi_1 = -\xi_3 = -\sqrt{13/15}$, $\xi_2 = 0$, and corresponding weights $W_1 = W_3 = 5/13$, $W_2 = 16/13$, in each natural coordinate direction (Challa, 1998).

D. Stabilized Galerkin methods

For low-order elements, reduced dispersion error may be achieved using residual-based methods such as Galerkin least-squares (GLS) and related methods. Least-squares stabilization stands out among the numerous approaches that have been proposed for reducing resolution requirements of standard Galerkin finite-element methods for time-harmonic acoustics by combining substantial improvement in accuracy with simple implementation. In the GLS method the Galerkin variational form is modified by appending residuals of the governing Helmholtz equation in a consistent least-squares form (Harari and Hughes, 1992),

$$B(w,u) + \int_{\tilde{\Omega}} \tau \mathcal{L}w \mathcal{L}u \, d\boldsymbol{x} = F(w). \tag{29}$$

Here, $\mathcal{L} = \nabla^2 + k^2$ is the Helmholtz differential operator, and $\tilde{\Omega}$ denotes integration over element interiors. The element parameter τ is usually determined from discrete dispersion

analysis and selected to minimize or eliminate dispersion error in the numerical solution.

For two-dimensional quadrilateral (Q4) elements, the value of τ is determined by enforcing numerical and continuous wave numbers to coincide $\tilde{k}=k$ (Thompson and Pinsky, 1995)

$$\tau=\frac{1}{k^2}\left[1-\frac{6}{(kh)^2}\left(\frac{1-\cos\alpha_x}{2+\cos\alpha_x}+\frac{1-\cos\alpha_y}{2+\cos\alpha_y}\right)\right], \tag{30}$$

where $(\alpha_x,\alpha_y)=kh(\cos\theta,\sin\theta)$. For uniform meshes this value eliminates dispersion error for plane waves in the angular direction θ. In general, solutions to the Helmholtz equation can be expanded in terms of plane waves with the predominant direction unknown a priori. By preselecting the angle $\theta_0=\pi/8$ in (30), the phase error, while not eliminated, is reduced significantly for all other wave angles (Thompson and Pinsky, 1995). For unstructured finite-element meshes the Laplacian operator appearing in (29) is usually neglected and the element size h can be taken as an average over the mesh or $h=\sqrt{A}$, where A is the element area. Numerical evidence shows that the GLS-FEM is relatively insensitive to the precise definition of the measure of element size (Harari and Magoules, 2004; Harari *et al.*, 1996; Thompson and Thangavelu, 2002). The additional cost of computing the GLS contribution is trivial, yet gives substantial improvement in accuracy, even on unstructured meshes. Values for τ on triangle, quadratic, and trilinear brick elements are given in Harari and Nogueira, 2002; Harari *et al.*, 1996; Thompson and Pinsky, 1995. Successful generalization of residual based methods to waves in plate bending elements and acoustic fluid—structure interaction are given in Thompson, 2003, and Thompson and Sankar, 2001.

Many of the generalized Galerkin methods can be derived within the variational multiscale (VMS) framework (Hughes *et al.*, 1998), including the method of residual-free bubbles (Franca *et al.*, 1997), also related to nearly optimal Petrov-Galerkin methods (Barbone and Harari, 2001). For the GLS method defined in (29), the mesh-dependent stability parameter τ may be interpreted as an algebraic approximation of a global integral operator for unresolved fine scales obtained by a separation of coarse finite-element polynomials and enhanced fine scales within the VMS framework. Multiscale considerations also underlie the residual-based method in Oberai and Pinsky, 2000, which includes the Helmholtz residual in least-squares form over element interiors $\tilde{\Omega}$, plus an additional residual defined over interelement boundaries $\tilde{\Gamma}$. In this case, the variational form assuming negligible Laplacian operation is

$$B(w,u)+k^2(\tau w,u)_{\tilde{\Omega}}-(\gamma w,[\![u_{,n}]\!])_{\tilde{\Gamma}}-(\gamma[\![w_{,n}]\!],u)_{\tilde{\Gamma}}=F(w),$$

where $[\![u_{,n}]\!]$ is the jump in discontinuous gradients across common element edges. Using this framework, Oberai and Pinsky (2000) find mesh parameters τ and γ for Q4 linear quadrilateral elements which produce a leading order phase error for all plane-wave directions of order $O(kh)^6$; a substantial improvement over standard Galerkin linear Q4 elements. Further discussions relating residual and other stabilized methods within the VMS framework are given in Harari, 2004.

E. Wave-based discretization methods

Wave-based methods exploit known solutions such as plane waves, or other analytical solutions to define or enrich the approximate solution space. Element-free methods (EFM) based on moving least-squares, and partition-of-unity methods (PUM), provide a means to incorporate analytical wave functions within local basis functions. For the Helmholtz equation solutions can be approximated using increasing numbers of basis functions in the form of plane waves (Babuska and Melenk, 1997; Bettess *et al.*, 2003; Laghrouche and Bettess, 2000; Laghrouche *et al.*, 2002; Melenk and Babuska, 1996); in 2D problems, $\mathcal{V}=\{e^{ik(x\cos\theta_m+y\sin\theta_m)},\theta_m=2\pi m/n\}$, where $m=0,1,\ldots,n-1$, $n=1,2,\ldots$. Suleau and Bouillard (2000) have shown that dispersion and pollution errors can be reduced by adding a sufficient number of plane-wave basis functions within the element-free moving least-squares method. Plane-wave basis functions have also been multiplied with standard piecewise polynomial shape functions as a partition-of-unity finite element method (Laghrouche and Bettess, 2000; Strouboulis *et al.*, 2000). The general integral that arises in the element matrices can be written as the product of plane waves, and polynomials. Such integrals are highly oscillatory and difficult to evaluate efficiently using standard techniques. Burnett and Soroka (1972) and Pierce *et al.* (2002) offer well-proven techniques for evaluating highly oscillatory integrals. Ortiz and Sanchez (2001) derive special integration techniques which isolate the oscillatory effects to one dimension. Another approach to reducing integration costs is given in Bettess *et al.*, 2003; Laghrouche *et al.*, 2002. While reducing dispersion and pollution error, a drawback of these approaches is the potential for ill-conditioning of the resulting system matrices which may disrupt the practical convergence of the method. Empirical rules relating condition number, number of wave directions, and wave number are given in Laghrouche *et al.*, 2002.

Least-squares methods (Monk and Wang, 1999; Stojek, 1998) minimize the least-squares difference in jumps of the solution and its normal derivative across element edges. The use of discontinuous solution spaces allows for the use of a plane-wave basis. The ultraweak variational formulation (Cessenat and Despres, 2003, 1998; Huttunen *et al.*, 2002) is another approach to using discontinuous local plane-wave solutions of the Helmholtz equation on each element. In this approach integration-by-parts is used to derive a variational formulation that weakly enforces continuity conditions between elements via local transmitting impedance conditions. An advantage of this approach is that integrations are carried out over element boundaries only and can be evaluated in closed form. Other wave-based methods are the weak element method (Goldstein, 1986), and the iterative defect-correction meshless method (Lacroix *et al.*, 2003). A difficulty with wave-based methods is that plane-wave, or other free-wave solutions, used as basis functions, often lead to ill-conditioning of the resulting system matrix as the number

of wave functions per element is increased, or refined element meshes are used. Another approach which may be derived in the framework of multiscale methods is the discontinuous enrichment method (DEM), where standard finite-element polynomial field is enriched within each element by adding plane-wave basis functions, and Lagrange multipliers are introduced at element interfaces to enforce a weak continuity of the solution (Farhat *et al.*, 2003). Using element level condensation, the system matrices are reported to be better conditioned than the PUM.

VIII. A POSTERIORI ERROR ESTIMATES AND ADAPTIVE METHODS

Adaptive methods use *a posteriori* error estimates to control discretization error in the numerical solution (Ainsworth and Oden, 2000). A posteriori error estimates are computed by postprocessing the numerical solution. For standard finite-element methods, a posteriori error estimates are used to control mesh refinement, both element size h, and polynomial order p distribution. The most common a posteriori error estimates are residual (both explicit and implicit), and recovery type. Implicit residual methods involve the solution of local or global problems which usually require very little cost compared to solving the original finite-element solution. Explicit methods relate the residual of the original governing equation localized to each element and do not require solving any auxiliary problems. Since the residual is a measure of numerical error, the error estimate can be used as a refinement indicator for adaptive strategies.

Stewart and Hughes (1996) used explicit residual methods based on the use of adjoint equations and duality arguments to develop adaptive strategies for the finite-element discretization of the Helmholtz equation. The error estimator depends on a stability constant which is approximated by solving global eigenvalue problems, which may be costly to compute. Irimie and Bouillard (2001) use standard explicit residual methods to estimate the error for the Helmholtz equation. Babuska *et al.* (1997) and Bouillard (1999) studied implicit element residual methods to estimate the finite-element error. Bouillard and Ihlenburg (1998, 1999) studied the gradient recovery-based error estimators based on the Zienkiewicz-Zhu patch recovery technique, and found that the estimator converges with mesh refinement for all wave numbers, although the estimate underestimates the error at high wave number. Attempts to estimate the pollution error are studied in Babuska *et al.* (1997). Initial studies of goal-oriented adaptive methods which measure the error in quantities of interest other than global norms are reported in Peraire and Patera, 1999; Sarrate *et al.*, 1999. For finite-element solutions with sufficient resolution and for which the pollution error is under control, i.e., $kL(kh)^2 \ll 1$, the quality of both the residual-based error estimators and recovery-type methods are good. However, since local error estimators do not detect pollution error, the quality deteriorates as the wave number k increases.

IX. ITERATIVE SOLUTION METHODS

The system of equations $\boldsymbol{Zd}=\boldsymbol{f}$, $\boldsymbol{Z} \in \mathbb{C}^{N_{\text{dof}} \times N_{\text{dof}}}$, resulting from the Galerkin FEM applied to the Helmholtz equation is sparse, complex, and symmetric (non-Hermitian and generally not diagonally dominant). Direct solution methods based on Gaussian elimination or factorization become exceedingly expensive both in terms of memory and computation when solving large systems of this class, especially for larger wave numbers. The primary iterative solution method for these types of systems are complex versions of Krylov subspace methods. In order to accelerate convergence a preconditioner is required. Standard preconditioners such as incomplete LU factorization (ILU) are optimal for matrices with diagonal dominance. However, the matrices which arise from the Helmholtz equation can be indefinite without diagonal dominance, making standard ILU not as practical and can exhibit breakdown at high wave numbers.

One possible approach is to recast the complex-symmetric system into Hermitian positive-definite form by multiplying by the Hermitian transpose, resulting in the normal equations $\boldsymbol{Z}^H\boldsymbol{Zd}=\boldsymbol{Z}^H\boldsymbol{f}$. The above system can be solved with the well-known conjugate gradient iterative method, with incomplete Cholesky factorization or other preconditioners. A difficulty of this approach is that the condition number of $\boldsymbol{Z}^H\boldsymbol{Z}$ is the square of $\boldsymbol{Z}$, so that convergence may be slow. Another possibility is to replace the complex-symmetric system $(\boldsymbol{A}+i\boldsymbol{B})(\boldsymbol{x}+i\boldsymbol{y})=\boldsymbol{a}+i\boldsymbol{b}$, with a real symmetric system,

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B} & -\boldsymbol{A} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ -\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix},$$

where $\boldsymbol{A} \in \mathbb{R}^{N_{\text{dof}} \times N_{\text{dof}}}$ and $\boldsymbol{B} \in \mathbb{R}^{N_{\text{dof}} \times N_{\text{dof}}}$. However, in this form, the system may become even more difficult to precondition.

Specialized preconditioners based on the original complex-symmetric matrix $\boldsymbol{Z}$ appear to give the best results with significant acceleration, even at high wave numbers. In Made (2001), prior to incomplete factorization, the real part of the preconditioning matrix is made less indefinite, by perturbations to the diagonal entries. This was shown to exhibit significant reduction of iteration counts in the GMRES method (Saad and Schultz, 1986). A similar idea is used in Erlangga *et al.*, 2004 with a shifted-Laplace preconditioner built on the modified Helmholtz equation $\nabla^2 - ik^2$, which is a generalization of the preconditioners introduced in Bayliss *et al.* (1983). In Kechroud *et al.* (2004) incomplete LU factorization with threshold (ILUT), and the Crout form of ILU were used as preconditioners combined with the GMRES iterative solver. It was reported that stagnation issues arise with high fill-in. The authors promise improvements by adding a damping parameter during the factorization process. Multifrontal incomplete factorization methods have also been applied to complex-symmetric indefinite systems with good results (Qu and Fish, 2002).

Suitable iterative solvers for sparse complex symmetric matrix systems other than GMRES include BICG-STAB (Vander Vorst, 1992), QMR (Freund, 1991) and TF-QMR (Freund, 1993) methods. Some numerical comparisons of the alterna-

tive iterative solvers are reported in Mazzia and Pini (2003) and elsewhere. The best combination of iterative solver and preconditioner seems to be problem dependent. From the author's experience, BICG-STAB appears to be robust for complex-symmetric systems at high wave number.

Other preconditioning approaches include analytic factorization methods (Gander and Nataf, 2001) and fictitious domain methods (FDM). The FDM is based on embedding the original domain into a larger one with a simple geometry (Marchuk *et al.*, 1986). In the algebraic FDM (closely related to capacitance matrix methods) the linear system resulting from the finite-element discretization is replaced by an equivalent, but enlarged, system corresponding to a simple-shaped domain containing the original domain. In this enlarged, yet simplified form, efficient iterative preconditioners are constructed based on locally perturbed orthogonal fitted meshes. Applications of the algebraic FDM for acoustic scattering are reported in Elman and O'Leary, 1998; Ernst, 1996; Heikkola *et al.*, 1999, 2003. Other solver techniques for the Helmholtz equation include multigrid methods (Brandt and Livshitz, 1997; Elman *et al.*, 2001; Lee *et al.*, 2000; Vanele *et al.*, 1998).

For acoustic scattering problems, the incident field is often represented by a plane wave, $u^{\text{inc}}(\boldsymbol{x})=\exp(ik\boldsymbol{x}\cdot\boldsymbol{\nu})$, $\boldsymbol{\nu}=(\cos\alpha,\sin\alpha\cos\beta,\sin\alpha\sin\beta)$ with a sweep over different incident directions α and β. This leads to a problem with fixed left-hand-side matrix and multiple right-hand-side forcing vectors. Malhotra *et al.* (1997) show how to efficiently solve the multiple right-hand-side problem with QMR methods.

A. Multifrequency solution methods

Often a large number of frequency (wave number) evaluations are required over a broad band to characterize the system response or when an inverse Fourier transform is needed to construct a corresponding time-domain solution. Since the complex-symmetric matrix $\boldsymbol{Z}=(\boldsymbol{S}-k^2\boldsymbol{M}-\boldsymbol{K}_\Gamma)$ is wave number dependent, the solution generally involves a separate inversion at each wave number k, causing the computation cost to grow linearly with the number of wave number evaluations. Djellouli *et al.* (2001) present an approach based on Padé approximation to compute multifrequency evaluations efficiently by solving a reference scattering problem with multiple excitation vectors and local BGT boundary conditions to characterize frequency derivatives of the scattered field. In Ingel (2000) and Thompson *et al.* (2001), domain decomposition concepts are combined with interpolation of substructure (subdomain) matrices over frequency bands of interest to accelerate multifrequency solutions.

For problems with frequency-independent excitation $\boldsymbol{f}$, matrix Padé approximation via the Lanczos (PVL) process can be used to obtain an efficient algorithm which grows sublinearly for the simultaneous solution of the Helmholtz equation over multiple frequencies in a window based on a Krylov projection to a subspace of much smaller dimension than the original system size (Feldmann and Freund, 1995; Freund, 1999; Simoncini and Perotti, 2002). Wagner *et al.* 2003a, b show that the PVL process can be implemented efficiently with the nonlocal modified DtN by utilizing the special structure of the admittance matrix. Efficient algorithms based on this PVL approach over multiple frequencies have not been demonstrated for the general case of frequency dependent excitation $\boldsymbol{f}(k)$ which is required for general acoustic scattering and radiation problems.

B. Domain decomposition methods

Domain decomposition methods provide an effective means of problem subdivision for parallel processing. Classical Schur complement-based domain-decomposition methods have difficulties when applied to the Helmholtz equation since the inversion of the matrix $\boldsymbol{A}_i=(\boldsymbol{S}_i-k^2\boldsymbol{M}_i)$ defined on each interior subdomain will be singular when the wave number corresponds to a resonance frequency (eigenvalue) of the pencil $(\boldsymbol{K}_i,\boldsymbol{M}_i)$. The first resonance will occur at a resolution of less than two subdomains per wavelength (Thompson *et al.*, 2001).

Kim (1998) uses the Robin-type impedance (transmission) interface conditions presented in Benamou and Despres (1997) and Despres (1993) in a Schwarz-type domain decomposition method to improve convergence; however, the iteration count increases with many subdomains. Cai *et al.* (1998) demonstrated an overlapping Schwarz method with GMRES acceleration and coarse grid corrections to improve convergence. Another additive Schwarz domain decomposition method with the Robin-type subdomain interface transmission conditions has been proposed in Susan-Resiga and Atassi (1998), where the nonlocal DtN nonreflecting boundary condition is computed with an iterative lag to maintain sparsity of the parallel subdomain solves. In Larsson (1999) a preconditioned restarted GMRES iterative method is used for the solving the Helmholtz equation, including nonlocal nonreflecting boundary conditions. Domain decomposition is used with a Schur complement algorithm and fast preconditioners for the subdomains to accelerate convergence.

In Farhat *et al.* (2000) and Tezaur *et al.* (2002), a nonoverlapping domain decomposition method called FETI-H, based on two-level Lagrange multipliers and the alternating Robin-type transmission conditions at subdomain interfaces of the continuous operator form $\mathcal{A}^{(s)}=ik$, presented in Benamou and Despres (1997) and Despres (1993), is used to solve the Helmholtz equation with second-order local nonreflecting boundaries. The matrices restricted to each subdomain Ω_s are $\boldsymbol{Z}^{(s)}=(\boldsymbol{S}^{(s)}-k^2\boldsymbol{M}^{(s)}-\boldsymbol{K}_\Gamma^{(s)}+ik\boldsymbol{A}^{(s)})$. Here, $\boldsymbol{A}^{(s)}$ are regularization matrices associated with the operator $\mathcal{A}^{(s)}=ik$, and defined by integration over subdomain interfaces,

$$\boldsymbol{d}^{(s)^T}\boldsymbol{A}^{(s)}\boldsymbol{d}^{(s)}=\sum_{\substack{t=1\\ t\neq s}}^{p}\epsilon^{st}\int_{\partial\Omega_s\cap\partial\Omega_t}u^2.$$

Here, $\boldsymbol{d}^{(s)}$ is the unknown vector restricted to subdomain Ω_s, and $\epsilon^{st}\in\{0,\pm1\}$, $\epsilon^{st}=-\epsilon^{ts}$. The use of regularization matrices for the Helmholtz operator provides for a unique solution on each subdomain. Inclusion of the alternating regularization matrix on the interface boundaries cancel upon global assembly, thus reverting to the original problem, and leading to nonsingular and invertible matrices $\boldsymbol{Z}^{(s)}$. The transmission

conditions can be interpreted as a simple local preconditioner of the linear system condensed on the interface. Improved transmission conditions with tangential derivatives of the form $\mathcal{A}^{(s)}=\alpha^{(s)}+\beta^{(s)}\partial^2_{\tau 2}$, with coefficients $\alpha^{(s)},\beta^{(s)}$, and unit tangential vector τ, have been derived based on Fourier analysis of the Steklov-Poincaré operators in a half-space (Collino *et al.*, 2000). Optimized coefficients have been chosen to minimize the convergence rate of the Jacobi algorithm in the closely related additive Schwarz method with no overlap (Gander *et al.*, 2002). In Magoules *et al.*, 2004, it is shown that the optimal augmented interface operator $\mathcal{A}^{(s)}$ is the Schur complement of the outer domain. Approximations of this Schur complement with sparse approximate inverse methods and incomplete factorization are investigated.

X. ACOUSTIC INVERSE PROBLEMS AND SHAPE OPTIMIZATION

Acoustic inverse problems which involve determining the shape of a body from an acoustic far-field pattern for the scattering of time-harmonic waves can be interpreted as a nonlinear ill-posed operator equation with the operator mapping the boundary onto the far field. Regularized Newton iterative methods provide accurate and robust solutions to the inverse obstacle scattering equation (Kress, 2003). Gradient-type methods based on regularized Newton or quasi-Newton optimization requires computation of shape derivatives. The derivatives are computed from solutions to the Helmholtz equation with different right-hand sides corresponding to the number of parameters in the boundary representation. Implementation of the regularized Newton method which incorporates exact sensitivities of far-field patterns with iterative finite-element solutions to the direct time-harmonic problem are given in Farhat *et al.* (2002). An alternate Newton-type method calculates the shape derivatives efficiently by solving an associated adjoint problem (Feijoo *et al.*, 2004). The adjoint method for the computation of shape derivatives and smoothing for finding regular solutions is related to the shape optimization problem (Jameson, 1988), where observations of a numerical solution to the Helmholtz equation provides an objective function which is formulated as a multipoint nonlinear optimization problem. Habbal (1998) and Bangtsson *et al.* (2003) consider shape optimization of a sound barrier and acoustic horn, respectively, using the regularized Newton iterations and the finite-element method. The shape optimization problem is solved using a gradient-based search method where the gradient is provided by solving the associated adjoint equations. Developments in coupled FEM-BEM structural-acoustic optimization and sensitivity analysis are reported in Fritze (2005) and Marburg (2002).

XI. CONCLUSION

Solutions to the challenging problem of developing accurate and efficient finite-element methods for time-harmonic acoustic radiation and scattering have a rich history of successes. It is expected that this trend will continue leading to many new and exciting developments in the future.

ACKNOWLEDGMENTS

I thank the anonymous reviewers for valuable comments and suggestions which greatly improved the quality of this article.

Ainsworth, M. (**2004**). "Discrete dispersion relation for *hp*-version finite element approximation at high wave number," SIAM (Soc. Ind. Appl. Math.) J. Numer. Anal. **42**(2), 553–575.

Ainsworth, M., and Oden, J. T. (**2000**). *A Posteriori Error Estimation in Finite Element Analysis* (Wiley, New York).

Antoine, X. (**2002**). "An algorithm coupling the OSRC and FEM for the computation of an approximate scattered acoustic field by a non-convex body," Int. J. Numer. Methods Eng. **54**(7), 1021–1041.

Antoine, X., Barucq, H., and Bendali, A. (**1999**). "Bayliss-Turkel-like radiation conditions on surfaces of arbitrary shape," J. Math. Anal. Appl. **229**(1), 184–211.

Astley, R. J. (**2000**). "Infinite elements for wave problems: A review of current formulations and an assessment of accuracy," Int. J. Numer. Methods Eng. **49**, 951–976.

Astley, R. J., and Coyette, J.-P. (**2001**). "The performance of spheroidal infinite elements," Int. J. Numer. Methods Eng. **52**(12), 1379–1396.

Astley, R. J., and Hamilton, J. A. (**2000**). "Numerical studies of conjugated infinite elements for acoustical radiation," J. Comput. Acoust. **8**(1), 1–24.

Astley, R. J., Macaulay, G. J., Coyette, J.-P., and Cremers, L. (**1998a**). "Three-dimensional wave-envelope elements of variable order for acoustic radiation and scattering. I. Formulation in the frequency domain," J. Acoust. Soc. Am. **103**, 49–63.

Astley, R. J., Macaulay, G. J., Coyette, J.-P., and Cremers, L. (**1998b**). "Three-dimensional wave-envelope elements of variable order for acoustic radiation and scattering. II. Formulation in the time domain," J. Acoust. Soc. Am. **103**, 64–72.

Atkinson, F. V. (**1949**). "On Sommerfeld's radiation condition," Philos. Mag. **40**, 645–651.

Babuska, I., and Melenk, J. M. (**1997**). "The partition of unity method," Int. J. Numer. Methods Eng. **40**, 727–758.

Babuska, I., Ihlenburg, F., Strouboulis, T., and Gangaraj, S. K. (**1997**). "A posteriori error estimation for finite element solutions of Helmholtz'equation. I. The quality of local indicators and estimators," Int. J. Numer. Methods Eng. **40**(18), 3443–3462.

Babuska, I., Ihlenburg, F., Strouboulis, T., and Gangaraj, S. K. (**1997**). "A posteriori error estimation for finite element solutions of Helmholtz' equation. II. Estimation of the pollution error," Int. J. Numer. Methods Eng. **40**(21), 3883–3900.

Bangtsson, E., Noreland, D., and Berggren, M. (**2003**). "Shape optimization of an acoustic horn," Comput. Methods Appl. Mech. Eng. **192**(11–12), 1533–1571.

Barbone, P. E., and Harari, I. (**2001**). "Nearly H^1-optimal finite element methods," Comput. Methods Appl. Mech. Eng. **190**, 5679–5690.

Bayliss, A., Goldstein, C. I., and Turkel, E. (**1983**). "An iterative method for Helmholtz equation," J. Comput. Phys. **49**, 443–457.

Bayliss, A., Gunzberger, M., and Turkel, E. (**1982**). "Boundary conditions for the numerical solution of elliptic equations in exterior domains," SIAM J. Appl. Math. **42**(2), 430–451.

Bayliss, A., Goldstein, C., and Turkel, E. (**1985**). "On accuracy conditions for the numerical computation of waves," J. Comput. Phys. **59**, 396–404.

Bayliss, A., Goldstein, C. I., and Turkel, E. (**1985**). "The numerical solution of the Helmholtz equation for wave propagation problems in underwater acoustics," Comput. Math. Appl. **11**(7/8), 655–665.

Benamou, J. D., and Despres, B. (**1997**). "A domain decomposition method for the Helmholtz equation and related optimal control problems," J. Comput. Phys. **136**, 68–82.

Berenger, J.-P. (**1994**). "A perfectly matched layer for the absorption of electromagnetic waves," J. Comput. Phys. **114**(2), 195–200.

Bettess, P. (**1977**). "Infinite elements," Int. J. Numer. Methods Eng. **11**(1), 53–64.

Bettess, P. (**1992**). *Infinite Elements* (Penshaw, Sunderland, UK).

Bettess, P., Shirron, J., Laghrouche, O., Peseux, B., Sugimoto, R., and Trevelyan, J. (**2003**). "A numerical integration scheme for special finite elements for the Helmholtz equation," Int. J. Numer. Methods Eng. **56**, 531–552.

Bossut, R., and Decarpigny, J.-N. (**1989**). "Finite element modeling of radiating structures using dipolar damping elements," J. Acoust. Soc. Am. **86**(4), 1234–1244.

Bouillard, P. (**1999**). "Influence of the pollution on the admissible field error estimation for FE solutions of the Helmholtz equation," Int. J. Numer. Methods Eng. **45**(7), 783–800.
Bouillard, P., and Ihlenburg, F. (**1998**). "Error estimation and adaptivity for the finite element method in acoustics," in *Advances in Adaptive Computational Methods in Mechanics*, edited by P. Ladeveze and J. T. Oden (Elsevier, Amsterdam), pp. 477–492.
Bouillard, P., and Ihlenburg, F. (**1999**). "Error estimation and adaptivity for the finite element method in acoustics: 2D and 3D applications," Comput. Methods Appl. Mech. Eng. **176**, 147–163.
Brandt, A., and Livshitz, I. (**1997**). "Ray-wave multigrid method for Helmholtz equation," Electron. Trans. Numer. Anal. **6**, 162–181.
Burnett, D. S. (**1994**). "A 3-D acoustic infinite element based on a generalized multipole expansion," J. Acoust. Soc. Am. **96**, 2798–2816.
Burnett, D. S., and Holford, R. L. (**1998a**). "Prolate and oblate spheroidal acoustic infinite elements," Comput. Methods Appl. Mech. Eng. **158**, 117–142.
Burnett, D. S., and Holford, R. L. (**1998b**). "An ellipsoidal acoustic infinite element," Comput. Methods Appl. Mech. Eng. **164**, 49–76.
Burnett, D. S., and Soroka, W. W. (**1972**). "An efficient numerical technique for evaluating large quantities of highly oscillatory integrals," Jpn. J. Appl. Phys. **10**, 325–332.
Burton, A., and Miller, G. (**1971**). "The application of integral equation methods to the numerical solution of some exterior boundary-value problems," Proc. R. Soc. London, Ser. A **323**, 201–210.
Cai, X. C., Cararin, M. A., Elliott, F. W., and Widlund, O. B. (**1998**). "Overlapping Schwarz algorithms for solving Helmholtz's equation," in *Domain Decomposition Methods 10*, edited by J. Mandel, C. Farhat, and X. C. Cai (Contemp. Math., AMS, Providence, RI), Vol. **218**, pp. 391–399.
Cessenat, O., and Despres, B. (**1998**). "Application of an ultra weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problem," SIAM (Soc. Ind. Appl. Math.) J. Numer. Anal. **35**(1), 255–299.
Cessenat, O., and Despres, B. (**2003**). "Using plane waves as base functions for solving time harmonic equations with the ultra weak variational formulation," J. Comput. Acoust. **11**(2), 227–238.
Challa, S. (1998). "High-order accurate spectral elements for wave propagation problems," Master's thesis, Clemson University, Mechanical Engineering Dept., August 1998.
Chen, J.-T., and Chen, K.-H. (**2004**). "Applications of the dual integral formulation in conjunction with fast multipole method in large-scale problems for 2D exterior acoustics," Eng. Anal. Boundary Elem. **28**, 685–709.
Chew, W. C., Jin, J. M., Lu, C. C., Michielssen, E., and Song, J. M. M. (**1997**). "Fast solution methods in electromagnetics," IEEE Trans. Antennas Propag. **45**(3), 533–543.
Chew, W. C., Song, J. M., Cui, T. J., Velarnparambil, S., Hastriter, M. L., and Hu, B. (**2004**). "Review of large scale computing in electromagnetics with fast integral equation solvers," Comput. Model. Eng. Sci. **5**(4), 361–372.
Collino, F., and Monk, P. B. (**1998**). "Optimizing the perfectly matched layer," Comput. Methods Appl. Mech. Eng. **164**(1–2), 157–171.
Collino, F., Ghanemi, S., and Joly, P. (**2000**). "Domain decomposition method for harmonic wave propagation: A general presentation," Comput. Methods Appl. Mech. Eng. **184**, 171–211.
Colton, D., and Kreiss, R. (**1983**). *Integral Equation Methods in Scattering Theory* (Wiley-Interscience, New York).
Craggs, A. (**1972**). "The use of simple three-dimensional acoustic finite elements for determining the natural modes and frequencies of complex shaped enclosures," J. Sound Vib. **23**(3), 331–339.
Darrigrand, E. (**2002**). "Coupling of fast multipole method and microlocal discretization for the 3-D Helmholtz equation," J. Comput. Phys. **181**(1), 126–154.
Darve, E. (**2000**). "The fast multipole method (i): Error analysis and asymptotic complexity," SIAM (Soc. Ind. Appl. Math.) J. Numer. Anal. **38**(1), 98–128.
Deraemaeker, A., Babuska, I., and Bouillard, P. (**1999**). "Dispersion and pollution of the FEM solution for the Helmholtz equation in one, two, and three dimensions," Int. J. Numer. Methods Eng. **46**, 471–499.
Despres, B. (**1993**). "Domain decomposition method and the Helmholtz problem (Part II)," in *2nd International Conference on Mathematical and Numerical Aspects of Wave Propagation*, Newark, DE (SIAM).
Dey, S. (**2003**). "Evaluation of p-FEM approximations for mid-frequency elasto-acoustics," J. Comput. Acoust. **11**(2), 195–225.
Djellouli, R., Farhat, C., and Tezaur, R. (**2001**). "A fast method for solving acoustic scattering problems in frequency bands," J. Comput. Phys. **168**, 412–432.
Dreyer, D., and von Estorff, O. (**2003**). "Improving conditioning of infinite elements for exterior acoustics," Int. J. Numer. Methods Eng. **58**(6), 933–953.
Elman, H. C., and O'Leary, D. P. (**1998**). "Efficient iterative solution of the three-dimensional Helmholtz equation," J. Comput. Phys. **142**, 163–181.
Elman, H. C., Ernst, O. G., and O'Leary, D. P. (**2001**). "A multigrid method enhanced by Krylov subspace iteration for discrete Helmholtz equations," SIAM J. Sci. Comput. (USA) **23**, 1290–1314.
Enquist, B., and Majda, A. (**1977**). "Absorbing boundary conditions for the numerical simulation of waves," Math. Comput. **31**(139), 629–651.
Epton, M. A., and Dembart, B. (**1995**). "Multipole translation theory for three-dimensional Laplace and Helmholtz equations," SIAM J. Sci. Comput. (USA) **16**(4), 865–897.
Erlangga, Y. A., Vuik, C., and Oosterlee, C. W. (**2004**). "On a class of preconditioners for solving the Helmholtz equation," Appl. Numer. Math. **50**(3–4), 409–425.
Ernst, O. G. (**1996**). "A finite-element capacitance matrix method for exterior Helmholtz problems," Numer. Math. **75**, 175–204.
Farhat, C., Harari, I., and Hetmaniuk, U. (**2003**). "The discontinuous enrichment method for multiscale analysis," Comput. Methods Appl. Mech. Eng. **192**, 3195–3209.
Farhat, C., Tezaur, R., and Djellouli, R. (**2002**). "On the solution of three-dimensional inverse obstacle acoustic scattering problems by a regularized Newton method," Inverse Probl. **18**(5), 1229–1246.
Farhat, C., Macedo, A., Lesoinne, M., Roux, F., Magoules, F., and La Bourdonnaie, A. (**2000**). "Two-level domain decomposition methods with Lagrange multipliers for the fast iterative solution of acoustic scattering problems," Comput. Methods Appl. Mech. Eng. **184**, 213–239.
Feijoo, G. R., Oberai, A. A., and Pinsky, P. M. (**2004**). "An application of shape optimization in the solution of inverse acoustic scattering problems," Inverse Probl. **20**(1), 199–228.
Feldmann, P., and Freund, R. W. (**1995**). "Efficient linear circuit analysis by Padé approximation via the Lanczos process," IEEE Trans. Comput.-Aided Des. **14**, 639–649.
Fischer, M., and Gaul, L. (**2005**). "Fast BEM-FEM mortar coupling for acoustic-structure interaction," Int. J. Numer. Methods Eng. **28**(62), 1677–1690.
Fischer, M., Gauger, U., and Gaul, L. (**2004**). "A multipole Galerkin boundary element method for acoustics," Eng. Anal. Boundary Elem. **28**(2), 155–162.
Franca, L. P., Farhat, C., Macedo, A. P., and Lesoinne, M. (**1997**). "Residual-free bubbles for the Helmholtz equation," Int. J. Numer. Methods Eng. **40**(21), 4003–4009.
Freund, R. W. (**1991**). "QMR: A quasi-minimal residual method of non-Hermitian linear systems," Numer. Math. **1**(60), 315–339.
Freund, R. W. (**1993**). "A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems," SIAM J. Sci. Comput. (USA) **14**(2), 470–482.
Freund, R. W. (**1999**). "Reduced-order modeling techniques based on Krylov subspaces and their use in circuit simulation," Applied and Computational Control, Signals, and Circuits **1**, 435–498.
Fritze, D., Marburg, S., and Hardtke, H.-J. (**2005**). "FEM-BEM coupling and structural acoustic sensitivity analysis for shell geometries," Comput. Struct. **83**, 143–154.
Gander, M. J., and Nataf, F. (**2001**). "AILU for Helmholtz problems: A new preconditioner based on the analytic parabolic factorization," J. Comput. Acoust. **9**(4), 1499–1509.
Gander, M. J., Magoules, F., and Nataf, F. (**2002**). "Optimized Schwarz methods without overlap for the Helmholtz equation," SIAM J. Sci. Comput. (USA) **24**, 38–60.
Gerdes, K. (**1998**). "Conjugated versus the unconjugated infinite element method for the Helmholtz equation in exterior domains," Comput. Methods Appl. Mech. Eng. **152**, 125–145.
Givoli, D. (**1999**). "Recent advances in the DtN FE method," Arch. Comput. Methods Eng. **6**(2), 71–116.
Givoli, D. (**2004**). "High-order local non-reflecting boundary conditions: A review," Wave Motion **39**, 319–326.
Givoli, D., Patlashenko, I., and Keller, J. B. (**1997**). "High order boundary conditions and finite elements for infinite domains," Comput. Methods Appl. Mech. Eng. **143**, 13–39.
Gladwell, G. M. L. (**1966**). "A variational formulation of damped acousto-structural vibration problems," J. Sound Vib. **4**, 172–186.
Goldstein, C. I. (**1986**). "The weak element method applied to Helmholtz

type equations," Appl. Numer. Math. **2**, 409–426.
Greengard, L., Huang, J., Rokhlin, V., and Stephen, W. (**1998**). "Accelerating fast multipole methods for the Helmholtz equation at low frequencies," IEEE Comput. Sci. Eng. **5**(3), 32–38.
Grote, M. J., and Keller, J. B. (**1995**). "On nonreflecting boundary conditions," J. Comput. Phys. **122**, 231–243.
Guddati, M. N., and Yue, B. (**2004**). "Modified integration rules for reducing dispersion error in finite element methods," Comput. Methods Appl. Mech. Eng. **193**(3–5), 275–287.
Gumerov, N. A., and Duraiswami, R. (**2004**). "Computation of scattering from clusters of spheres using the fast multipole method," J. Acoust. Soc. Am. **117**, 1744–1761.
Habbal, A. (**1998**). "Nonsmooth shape optimization applied to linear acoustics," SIAM J. Optim. **8**(4), 989–1006.
Hagstrom, T. (**1999**). *Radiation Boundary Conditions for the Numerical Simulation of Waves* (Cambridge University Press, Cambridge), Vol. **8**.
Hagstrom, T., and Hariharan, S. I. (**1998**). "A formulation of asymptotic and exact boundary conditions using local operators," Appl. Numer. Math. **27**, 403–416.
Harari, I. (**2004**). "Acoustics," in *Finite Element Methods: 1970's and Beyond*, edited by L. P. Franca, T. E. Tezduyar, and A. Masud (CIMNE, Barcelona).
Harari, I., and Hughes, T. J. R. (**1992**). "A cost comparison of boundary element and finite element methods for problems of time-harmonic acoustics," Comput. Methods Appl. Mech. Eng. **97**, 77–102.
Harari, I., and Hughes, T. J. R. (**1992**). "Analysis of continuous formulations underlying the computation of time-harmonic acoustics in exterior domains," Comput. Methods Appl. Mech. Eng. **97**, 103–124.
Harari, I., and Hughes, T. J. R. (**1992**). "Galerkin/least-squares finite element methods for the reduced wave equation with non-reflecting boundary conditions in unbounded domains," Comput. Methods Appl. Mech. Eng. **98**, 411–454.
Harari, I., and Magoules, F. (**2004**). "Numerical investigations of stabilized finite element computations for acoustics," Wave Motion **39**(4), 339–349.
Harari, I., and Nogueira, C. L. (**2002**). "Reducing dispersion of linear triangular elements for the Helmholtz equation," J. Eng. Mech. **128**(3), 351–358.
Harari, I., Slavutin, M., and Turkel, E. (**2000**). "Analytical and numerical studies of a finite element PML for the Helmholtz equation," J. Comput. Acoust. **8**(1), 121–137.
Harari, I., Grosh, K., Hughes, T. J. R., Malhotra, M., Pinsky, P. M., Stewart, J. R., and Thompson, L. L. (**1996**). "Recent developments in finite element methods for structural acoustics," Arch. Comput. Methods Eng. **3**, 132–311.
Heikkola, E., Kuznetsov, Y. A., and Lipnikov, K. (**1999**). "Fictitious domain methods for the numerical solution of three-dimensional acoustic scattering problems," J. Comput. Acoust. **7**, 161–183.
Heikkola, E., Rossi, T., and Toivanen, J. (**2003**). "A parallel fictitious domain method for the three-dimensional Helmholtz equation," SIAM J. Sci. Comput. (USA) **24**(5), 1567–1588.
Huan, R., and Thompson, L. L. (**2000**). "Accurate radiation boundary conditions for the time-dependent wave equation on unbounded domains," Int. J. Numer. Methods Eng. **47**, 1569–1603.
Hughes, T. J. R. (**2000**). *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis* (Dover Publications, New York).
Hughes, T. J. R., Feijoo, G. R., Mazzei, L., and Quincy, J.-B. (**1998**). "The variational multiscale method—a paradigm for computational mechanics," Comput. Methods Appl. Mech. Eng. **166**, 3–24.
Hunt, J. T., Knittel, M. R., and Barach, D. (**1974**). "Finite-element approach to acoustic radiation from elastic structures," J. Acoust. Soc. Am. **55**(2), 269–280.
Hunt, J. T., Knittel, M. R., Nichols, C. S., and Barach, D. (**1975**). "Finite-element approach to acoustic scattering from elastic structures," J. Acoust. Soc. Am. **57**(2), 287–299.
Huttunen, T., Monk, P., and Kaipio, J. (**2002**). "Computational aspects of the ultra-weak variational formulation," J. Comput. Phys. **182**(1), 27–46.
Ianculescu, C., and Thompson, L. L. (**2003**). "Parallel iterative finite element solution methods for three-dimensional acoustic scattering," 2003 ASME International Mechanical Engineering Congress & Exposition, Washington, D.C., 16–21, Nov. 2003, Paper IMECE2003-44266.
Ianculescu, C., and Thompson, L. L. (in press). "Parallel iterative solution for the Helmholtz equation with exact non-reflecting boundary conditions," Comput. Methods Appl. Mech. Eng..
Ihlenburg, F. (**1998**). *Finite Element Analysis of Acoustic Scattering* (Springer, Berlin).
Ihlenburg, F. (**2000**). "On fundamental aspects of exterior approximations with infinite elements," J. Comput. Acoust. **8**(1), 63–80.
Ihlenburg, F. (**2003**). "The medium-frequency range in computational acoustics: Practical and numerical aspects," J. Comput. Acoust. **11**(2), 175–194.
Ihlenburg, F., and Babuska, I. (**1995**). "Dispersion analysis and error estimation of Galerkin finite element methods for the Helmholtz equation," Int. J. Numer. Methods Eng. **38**, 3745–3774.
Ingel, R. P., Dyka, C. T., and Flippen, L. D. (**2000**). "Model reduction and frequency windowing for acoustic FEM analysis," J. Sound Vib. **238**(2), 327–350.
Irimie, S., and Bouillard, P. (**2001**). "A residual a posteriori error estimator for the finite element solution of the Helmholtz equation," Comput. Methods Appl. Mech. Eng. **190**(31), 4027–4042.
Jameson, A. (**1988**). "Aerodynamic design via control theory," J. Scientific Computing **3**, 233–260.
Kang, F. (**1983**). "Finite element method and natural boundary reduction," in Proceedings of the International Congress of Mathematicians, pp. 1439–1453. Warsaw.
Kechroud, R., Soulaimani, A., Saad, Y., and Gowda, S. (**2004**). "Preconditioning techniques for the solution of the Helmholtz equation by the finite element method," Math. Comput. Simul. **65**, 303–321.
Keller, J. B., and Givoli, D. (**1989**). "Exact non-reflecting boundary conditions," J. Comput. Phys. **81**, 172–192.
Kim, S. (**1998**). "Domain decomposition iterative procedures for solving scalar waves in the frequency domain," Numer. Math. **79**, 231–259.
Kress, R. (**2003**). "Newton's method for inverse obstacle scattering meets the method of least-squares," Inverse Probl. **19**(6), S91–S104.
Lacroix, V., Bouillard, Ph., and Villon, P. (**2003**). "An iterative defect-correction type meshless method for acoustics," Int. J. Numer. Methods Eng. **57**, 2131–2146.
Laghrouche, O., and Bettess, P. (**2000**). "Short wave modelling using special finite elements," J. Comput. Acoust. **8**(1), 189–210.
Laghrouche, O., Bettess, P., and Astley, R. J. (**2002**). "Modeling of short wave diffraction problems using approximating systems of plane waves," Int. J. Numer. Methods Eng. **54**, 1501–1533.
Larsson, E. (**1999**). "A domain decomposition method for the Helmholtz equation in a multilayer domain," SIAM J. Sci. Comput. (USA) **20**(5), 1713–1731.
Lee, B., Manteuffel, T. A., McCormick, S. F., and Ruge, J. (**2000**). "First-order system least-squares for the Helmholtz equation," SIAM J. Sci. Comput. (USA) **21**, 1927–1949.
Leis, R. (**1986**). *Initial-boundary Value Problems in Mathematical Physics* (Teubner, Stuttgart).
Made, M. M. M. (**2001**). "Incomplete factorization-based preconditionings for solving the Helmholtz equation," Int. J. Numer. Methods Eng. **50**, 1077–1101.
Magoules, F., Roux, F.-X., and Salmon, S.(**2004**). "Optimal discrete transmission conditions for a nonoverlapping domain decomposition method for the Helmholtz equation," SIAM J. Sci. Comput. (USA) **25**(5), 1497–1515.
Malhotra, M., and Pinsky, P. M. (**1996**). "A matrix-free interpretation of the nonlocal Dirichlet-to-Neumann radiation boundary condition," Int. J. Numer. Methods Eng. **39**, 3705–3713.
Malhotra, M., Freund, R. W., and Pinsky, P. M. (**1997**). "Iterative solution of multiple radiation and scattering problems in structural acoustics using a block quasi-minimal residual algorithm," Comput. Methods Appl. Mech. Eng. **146**, 173–196.
Marburg, S. (**2002**). "Developments in structural-acoustic optimization for passive noise control," Arch. Comput. Methods Eng. **9**(4), 291–370.
Marchuk, G. I., Kuznetsov, Y. A., and Matsokin, A. M. (**1986**). "Fictitious domain and domain decomposition methods," Sov. J. Numer. Anal. Math. Modelling **1**, 3–35.
Mazzia, A., and Pini, G. (**2003**). "Numerical performance of preconditioning techniques for the solution of complex sparse linear systems," Commun. Numer. Methods Eng. **19**, 37–49.
Melenk, J. M., and Babuska, I. (**1996**). "The partition of unity method: basic theory and applications," Comput. Methods Appl. Mech. Eng. **139**, 289–314.
Monk, P., and Wang, D.-Q. (**1999**). "A least-squares method for the Helmholtz equation," Comput. Methods Appl. Mech. Eng. **175**, 121–136.
Murphy, J. E., and Ching-Bing, S. A. (**1989**). "A finite-element model for ocean acoustic propagation and scattering," J. Acoust. Soc. Am. **86**(4),

1478–1483.
Nefske, D. J., Wolf, J. A., and Howell, L. J. (**1982**). "Structural-acoustic finite element analysis of the automobile passenger compartment: A review of current practice," J. Sound Vib. **80**(2), 247–266.
Oberai, A. A., and Pinsky, P. M. (**2000**). "A residual-based finite element method for the Helmholtz equation," Int. J. Numer. Methods Eng. **49**, 399–419.
Oberai, A. A., Malhotra, M., and Pinsky, P. M. (**1998**). "On the implementation of the Dirichlet-to-Neumann radiation condition for iterative solution of the Helmholtz equation," Appl. Numer. Math. **27**, 443–464.
Ortiz, P., and Sanchez, E. (**2001**). "An improved partition of unity finite element model for diffraction problems," Int. J. Numer. Methods Eng. **50**, 2727–2740.
Peraire, J., and Patera, A. T. (**1999**). "Asymptotic a posteriori finite element bounds for the outputs of noncoercive problems: The Helmholtz and Burgers equations," Comput. Methods Appl. Mech. Eng. **171**(1), 77–86.
Petyt, M. (**1990**). *Introduction to Finite Element Vibration Analysis* (Cambridge University Press, Cambridge).
Petyt, M., Lea, J., and Koopmann, G. H. (**1976**). "A finite element method for determining the acoustic modes of irregular shaped cavities," J. Sound Vib. **45**(4), 495–502.
Pierce, A. D., Cleveland, R. O., and Zampolli, M. (**2002**). "Radiation impedance matrices for rectangular interfaces within rigid baffles: Calculation methodology and applications," J. Acoust. Soc. Am. **111**(2), 672–684.
Qu, Y., and Fish, J. (**2002**). "Multifrontal incomplete factorization for indefinite and complex symmetric systems," Int. J. Numer. Methods Eng. **53**(6), 1433–1459.
Rockhlin, V. (**1993**). "Diagonal forms of the translation operators for the Helmholtz equation in three dimensions," Appl. Comput. Harmon. Anal. **1**, 82–98.
Saad, Y., and Schultz, M. H. (**1986**). "GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems," SIAM J. Sci. Comput. (USA) **7**, 856–869.
Sakuma, T., and Yasuda, Y. (**2002**). "Fast multipole boundary element method for large scale steady state sound field analysis. I. Setup and validation," Acust. Acta Acust. 88, 513–525.
Sarrate, J., Peraire, J., and Patera, A. T. (**1999**). "A posteriori finite element error bounds for non-linear outputs of the Helmholtz equation," Int. J. Numer. Methods Fluids **31**(1), 17–36.
Schneider, S. (**2003**). "Application of fast methods for acoustic scattering and radiation problems," J. Comput. Acoust. **11**(3), 387–401.
Shirron, J. J., and Babuska, I. (**1998**). "A comparison of approximate boundary conditions and infinite element methods for exterior Helmholtz problems," Comput. Methods Appl. Mech. Eng. **164**(1–2), 121–139.
Shirron, J. J., and Dey, S. (**2002**). "Acoustic infinite elements for non-separable geometries," Comput. Methods Appl. Mech. Eng. **191**, 4123–4139.
Simoncini, V., and Perotti, F. (**2002**). "On the numerical solutions of $(\lambda^2 a + \lambda b + c)x = b$ and application to structural dynamics," SIAM J. Sci. Comput. (USA) **23**(6), 1875–1897.
Sommerfeld, A. (**1912**). "Die Greensche Funktion der Schwingungsgleichung," Jahresber. Deutsch. Math. **21**, 309–353.
Stewart, J. R., and Hughes, T. J. R. (**1996**). "Explicit residual-based a posteriori error estimation for finite element discretizations of the Helmholtz equation. Computation of the constant and new measures of error estimator quality," Comput. Methods Appl. Mech. Eng. **131**(3–4), 335–363.
Stojek, M. (**1998**). "Least-squares Trefftz-type elements for the Helmholtz equation," Int. J. Numer. Methods Eng. **41**, 831–849.
Strouboulis, T., Babuska, I., and Copps, K. (**2000**). "The design and analysis of the generalized finite element method," Comput. Methods Appl. Mech. Eng. **181**(1–3), 43–69.
Suleau, S., and Bouillard, P. (**2000**). "Dispersion and pollution of meshless solutions for the Helmholtz equation," Comput. Methods Appl. Mech. Eng. **190**, 639–657.
Susan-Resiga, R. F., and Atassi, H. M. (**1998**). "A domain decomposition method for the exterior Helmholtz problem," J. Comput. Phys. **147**, 388–401.
Tezaur, R., Macedo, A., and Farhat, C. (**2001**). "Iterative solution of large-scale acoustic scattering problems with multiple right hand-sides by a domain decomposition method with lagrange multipliers," Int. J. Numer. Methods Eng. **51**(10), 1175–1193.
Tezaur, R., Macedo, A., Farhat, C., and Djellouli, R. (**2002**). "Three-dimensional finite element calculations in acoustic scattering using arbitrarily shaped convex artificial boundaries," Int. J. Numer. Methods Eng. **53**(6), 1461–1476.
Thompson, L. L. (**2003**). "On optimal stabilized MITC4 plate bending elements for accurate frequency response analysis," Comput. Struct. **81**, 995–1008.
Thompson, L. L., and Kunthong, P. (**2005**). "A residual based variational method for reducing dispersion error in finite element methods," *2005 ASME International Mechanical Engineering Congress and Exposition*. 5–11, November 2005, Orlando, Florida. The American Society of Mechanical Engineers, Paper IMECE2005-80551.
Thompson, L. L., and Pinsky, P. M. (**1994**). "Complex wavenumber Fourier analysis of the p-version finite element method," Comput. Mech. **13**(4), 255–275.
Thompson, L. L., and Pinsky, P. M. (**1995**). "A Galerkin least squares finite element method for the two-dimensional Helmholtz equation," Int. J. Numer. Methods Eng. **38**, 371–397.
Thompson, L. L., and Pinsky, P. M. (**1996**). "A space-time finite element method for structural acoustics in infinite domains. II. Exact time-dependent non-reflecting boundary conditions," Comput. Methods Appl. Mech. Eng. **132**, 229–258.
Thompson, L. L., and Pinsky, P. M. (**2004**). "Acoustics," in *Encyclopedia of Computational Mechanics*, edited by E. Stein, R. De Borst, and J. R. Hughes (Wiley InterScience, New York), Vol. **2**, Chap. 22.
Thompson, L. L., and Sankar, S. (**2001**). "Dispersion analysis of stabilized finite element methods for acoustic fluid interaction with Reissner-Mindlin plates," Int. J. Numer. Methods Eng. **50**(11), 2521–2545.
Thompson, L. L., and Thangavelu, S. R. (**2002**). "A stabilized MITC element for accurate wave response in Reissner-Mindlin plates," Comput. Struct. **80**(9–10), 769–789.
Thompson, L. L., Huan, R., and Ianculescu, C. (**1999**). "Finite element formulation of exact Dirichlet-to-Neumann radiation conditions on elliptic and spheroidal boundaries," in *1999 ASME International Mechanical Engineering Congress & Exposition*, Nashville, TN, 14–19, Nov. 1999, ASME Noise Control and Acoustics Division—1999, NCA-Vol. **26**, pp. 497–510.
Thompson, L. L., Huan, R., and He, D. (**2001**). "Accurate radiation boundary conditions for the two-dimensional wave equation on unbounded domains," Comput. Methods Appl. Mech. Eng. **191**, 311–351.
Thompson, L. L., Zhang, L., and Ingel, R. P. (**2001**). "Domain decomposition methods with frequency band interpolation for computational acoustics," in *2001 ASME International Mechanical Engineering Congress and Exposition*, 11–16, November 2001, New York, New York. The American Society of Mechanical Engineers, IMECE2001, Paper NCA-23532.
Tsynkov, S. V., and Turkel, E. (**2001**). "A Cartesian perfectly matched layer for the Helmholtz equation," in *Absorbing Boundaries and Layers, Domain Decomposition Methods*, edited by L. Tourrette and L. Halpern (Nova Science, New York), pp. 279–309.
Turkel, E., and Yefet, A. (**1998**). "Absorbing PML boundary layers for wave-like equations," Appl. Numer. Math. **27**(4), 533–557.
Van der Vorst, H. A. (**1992**). "BI-CGSTAB: A fast and smoothly converging variant of BI-CG for the solution of nonsymmetric linear systems," SIAM J. Sci. Comput. (USA) **13**(2), 631–644.
Van Hal, B., Desmet, W., Vandepitte, D., and Sas, P. (**2003**). "A coupled finite element-wave based approach for the steady-state dynamic analysis of acoustic systems," J. Comput. Acoust. **11**(2), 285–303.
van Joolen, V., Givoli, D., and Neta, B. (**2003**). "High-order non-reflecting boundary conditions for dispersive waves in cartesian, cylindrical and spherical coordinate systems," Int. J. Comput. Fluid Dyn. **17**(4), 263–274.
Vanek, P., Mandel, J., and Brezina, M. (**1998**). "Two-level algebraic multigrid for the Helmholtz problem," Contemp. Math. **218**, 349–356.
Wagner, M. M., Pinsky, P. M., and Malhotra, M. (**2003a**). "Application of Padé via Lanczos approximations for efficient multifrequency solution of Helmholtz problems," J. Acoust. Soc. Am. **113**(1), 313–319.
Wagner, M. M., Pinsky, P. M., Oberai, A. A., and Malhotra, M. (**2003b**). "A Krylov subspace projection method for simultaneous solution of Helmholtz problems at multiple frequencies," Comput. Methods Appl. Mech. Eng. **192**, 4609–4640.
Walsh, T., Demkowicz, L., and Charles, R. (**2004**). "Boundary element modeling of the external human auditory system," J. Acoust. Soc. Am. **115**(3), 1033–1043.
Wilcox, C. H. (**1956**). "An expansion theorem for electromagnetic fields," Commun. Pure Appl. Math. **9**, 115–134.
Wu, J.-Y., Kingsland, D. M., Lee, J.-F., and Lee, R. (**1997**). "A comparison of anisotropic PML to Berenger's PML and its application to the finite-element method for EM scattering," IEEE Trans. Antennas Propag. **45**(1),

40–50.
Young, J. C.-I., and Crocker, M. J. (**1975**). "Prediction of transmission loss in mufflers by the finite element method," J. Acoust. Soc. Am. **57**(1), 144–148.
Zienkiewicz, O. C. (**2000**). "Achievements and some unsolved problems of the finite element method," Int. J. Numer. Methods Eng. **47**(1–3), 9–28.
Zienkiewicz, O. C., and Newton, R. E. (**1969**). "Coupled vibrations of a structure submerged in a compressible fluid," in Proc. Int. Symp. Finite Element Techniques, Stuttgart, Germany, 1969. Int. Assoc. Ship Structures.

Effect of masker-fringe onset asynchrony on overshoot (L)[a)]

Andrzej Miśkiewicz[b)]
Institute of Hearing, Speech & Language and Department of Speech-Language Pathology & Audiology (106A FR), Northeastern University, Boston, Massachusetts 02115

Søren Buus
Institute of Hearing, Speech & Language and Communications & Digital Signal Processing Center, Electrical & Computer Engineering Department (440 DA), Northeastern University, Boston, Massachusetts, 02115

Mary Florentine[c)]
Institute of Hearing, Speech & Language and Department of Speech-Language Pathology & Audiology (106A FR), Northeastern University, Boston, Massachusetts 02115

(Received 24 July 2003; revised 2 November 2005; accepted 13 November 2005)

The study examines how overshoot is influenced by masker-signal onset asynchrony when the masker contains frequencies above or below the signal frequency. Masked thresholds were measured for a 2-ms tone at 5 kHz. The measurements were made in a reference condition with a narrow center-band (CB) noise masker (4590–5464 Hz), and in conditions with either a low-fringe (1900–4590 Hz) or a high-fringe noise band (5500–11 000 Hz) added to the CB. The signal always came on 2 ms after the onset of the CB. The time interval between fringe and signal onsets varied from −98 ms (signal onset before fringe onset) to +502 ms (signal onset after fringe onset). Results show that overshoot is largest, 8–11 dB, for a high fringe with onset occurring between 8 ms before the signal onset and 12 ms after it. Thus, pronounced overshoot is observed even when the high fringe is gated on *after* the signal's offset. Low fringes produce no more than 4 dB of overshoot, much less than high fringes produce. The finding of pronounced overshoot elicited by a high fringe presented shortly after the end of the signal suggests that overshoot is governed, at least in part, by central mechanisms. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2162175]

PACS number(s): 43.66.Dc, 43.66.Mk [NFV] Pages: 1331–1334

I. INTRODUCTION

The term "overshoot," introduced by Zwicker (1965a), refers to the finding that the threshold for a brief tone is considerably higher when the tone is presented within a few milliseconds after the onset of a noise masker than when it is presented with a delay greater or equal to 100 ms. Overshoot depends on various factors, such as the signal frequency and duration, the masker-signal onset delay, and the masker level and frequency content [see Bacon and Smith (1991) and Strickland (2001) for reviews].

Overshoot does not occur unless the masker contains energy outside the auditory filter centered at the signal frequency (e.g., Zwicker, 1965b). Schmidt and Zwicker (1991) showed that overshoot is elicited predominantly by maskers above the signal frequency, whereas masker bands below the signal frequency produce very little, if any, overshoot. The asymmetrical effects of masker spectrum were also observed by authors (Bacon and Smith, 1991; Carlyon and White, 1992) who examined how overshoot changes with the duration of a temporal gap introduced in various frequency bands of an otherwise continuous noise masker, and by authors (e.g., Hicks and Bacon, 1992) who investigated the reduction of overshoot caused by a preceding noise burst gated off shortly before the masker's onset.

The present study was carried out to determine how overshoot depends on the masker-signal onset asynchrony when a fringe masker band is located above the signal frequency and when it is located below. The measurements were made for a wide range of time intervals between the masker and signal onsets and included conditions with signal onset not only following but also preceding the fringe masker onset.

II. METHOD

A. Stimuli and apparatus

Signal frequency and duration as well as masker spectra and levels were similar to those used by Schmidt and Zwicker (1991). Masked thresholds were measured for 2-ms tone bursts at 5 kHz. Each tone burst consisted of a 1-ms raised-cosine rise followed immediately by a fall. The

[a)]Portions of this study were presented at the 147th Meeting of the Acoustical Society of America [J. Acoust. Soc. Am.**115**, 2602(A) (2004)].

[b)]Current affiliation: Musical Acoustics Laboratory, Department of Sound Engineering, The Frederic Chopin Academy of Music, Okólnik 2, 00-368 Warsaw, Poland; electronic mail: misk@chopin.edu.pl

[c)]Electronic mail: florentin@neu.edu

measurements were made for three masker bands, filtered from pink noise: (a) a narrow center-band masker (CB, 4590–5464 Hz), which was centered on the signal frequency, (b) a low fringe (LF, 1900–4590 Hz), and (c) a high fringe (HF, 5500–11 000 Hz). The bandwidths, expressed in critical-band-rate units, were 1 bark (CB), 5.3 barks (LF), and 3.7 barks (HF). The overall sound pressure level of pink noise was 60 dB and the SPLs of the maskers depended on their bandwidths. The SPLs of the masker bands were 46 dB (CB), 53 dB (LF), and 52 dB (HF). Thresholds were measured in a control condition with the CB alone and in the experimental conditions with the CB plus one of the fringes. The CB, too narrow to cause significant overshoot (Zwicker, 1965b), provided a stable measure of the masked threshold without a fringe; any alteration in that threshold would result from the overshoot mechanism controlled by masker energy outside the auditory filter centered at the signal frequency.

Figure 1 shows the temporal sequences of the signal and the two masker bands. With respect to the CB masker, the signal was always gated on 2 ms after masker onset. In the experimental conditions, with respect to the fringe masker, the signal onset either followed the onset of the fringe, was simultaneous with fringe onset, or preceded the onset of the fringe. The CB and the fringe were gated off simultaneously; they lasted at least 30 ms, but were lengthened as necessary to allow them to be turned off simultaneously.

The experiment comprised three parts covering different ranges of delays between the onsets of the fringe and the signal, as shown in Table I. Some conditions were repeated in different parts of the experiment to check the consistency of each listener's threshold. Part 2 also included a stimulus condition with a continuous fringe added to the gated CB. In each part, the experimental conditions were grouped in two sets of runs shown in separate rows of Table I. The order of conditions within a set of runs was randomized, as was the order of sets in each part.

Testing was carried out in a sound-attenuating booth. Stimulus generation and presentation were controlled by a PDP 11/03 computer. The signal was generated by a Krohn-Hite 4180 programmable oscillator. The CB masker was obtained from a BK 1405 pink-noise generator and a digital FIR filter (IHR Universal Filter); the fringes were obtained from a GR 1382 noise generator and a brickwall filter (Rockland 753, 110 dB/octave). The signal and the two noise bands were independently gated (Wilsonics BSIT) and attenuated (Wilsonics PATT) before being summed, amplified, and sent to one earphone of a Yamaha YH-1 headphone set. The SPLs reported in the following are based on a calibration level of 92 dB SPL for a 1-V input.

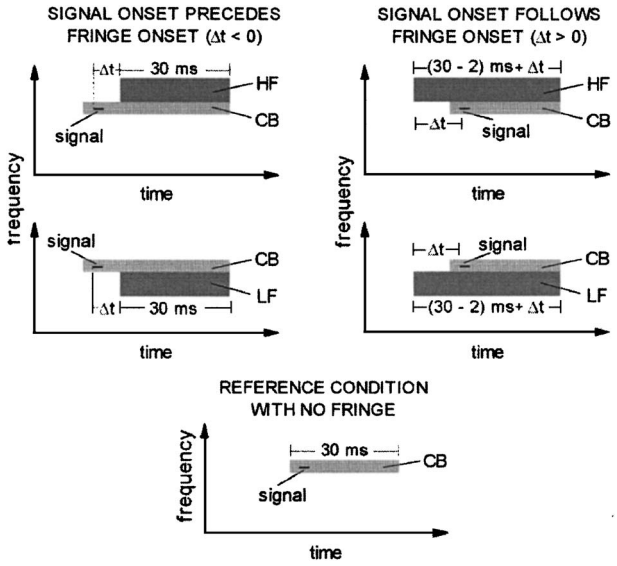


FIG. 1. Schematic showing the temporal sequences and the spectra of stimuli: CB—center-band masker (4590–5464 Hz), HF—high fringe (5500–11 000 Hz), LF—low fringe (1900–4590 Hz).

B. Procedure and listeners

Thresholds were measured by an adaptive, two-interval, two-alternative forced-choice procedure with visual feedback. The masker burst was presented in both observation intervals and the signal occurred with equal *a priori* probability in either the first or the second interval. The interval between the onsets of the CB masker in the two observation intervals was 320 ms in parts 1 and 3 of the experiment and 1000 ms in part 2.

The signal level varied according to a three-down/one-up decision rule (Levitt, 1971). A single threshold estimate was based on three consecutive tracks. Each track started with the signal at 76 dB SPL in LF and CB-alone conditions, and at 91 dB SPL in HF conditions. The initial step size of 5 dB was reduced to 2 dB after the first reversal of signal level. The track terminated after five reversals, and the average signal level at the last two reversals was taken as the threshold value for the track. The thresholds from the three consecutive tracks were averaged to provide one overall threshold estimate. Five such estimates from a total of 15 brief tracks were obtained for each listener and condition.

TABLE I. Summary of signal delays and masker conditions in parts 1–3 of the experiment. Under CB only the center masker was present with the signal delayed 2 ms; under Cnt the fringe was continuous. The numerical values are delays in milliseconds between the onsets of the fringe-band masker and the signal. A negative delay indicates that the signal onset preceded the fringe onset.

		CB	−98	−48	−18	−8	−2	0	2	4	6	12	22	52	102	202	502	Cnt
Part 1	LF	×				×	×	×	×	×	×	×						
	HF					×	×	×	×	×	×	×						
Part 2	LF	×							×			×			×	×	×	×
	HF	×							×			×			×	×	×	×
Part 3	LF, HF	×										×	×	×	×			
	LF, HF	×	×	×	×	×												

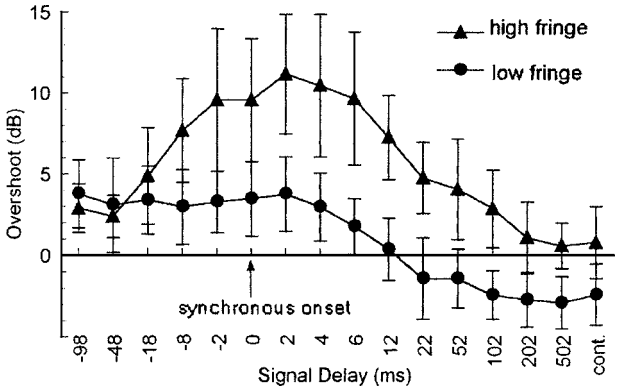


FIG. 2. Average amount of overshoot from seven listeners as a function of signal onset delay, relative to the onset of the fringe band. Data obtained with the HF (triangles) and LF (circles) added to the CB masker. Negative values on the abscissa indicate that the signal was gated on before the fringe's onset. Error bars show the standard deviations. The average threshold obtained with CB along was 61.4 dB SPL.

Data were obtained from seven listeners, six of whom were female. Six listeners were students, aged 18–28, and one was a member of the staff, 39 years old. All listeners had normal hearing. Their audiometric thresholds were 10 dB HL or less at octave frequencies between 0.25 and 8 kHz (ANSI, 1989). At the beginning of the study, each listener participated in a practice session that included a complete set of the conditions used in part 1 of the experiment.

C. Data analysis

Because no consistent differences were observed among measurements that were repeated in the three parts of the experiment, repetitions were combined. In most conditions, an individual threshold is the mean of five measurements. For conditions that were repeated in different parts of the experiment (see Table I), the individual thresholds are based on more than five measurements. Following Schmidt and Zwicker (1991), the amount of overshoot was calculated as the difference in decibels between the threshold measured with a fringe added to the CB and that measured with only the CB masker.

III. RESULTS

Figure 2 presents the mean amount of overshoot, averaged across the seven listeners, as a function of fringe-signal onset delay. At most delays, overshoot is much larger with a high fringe than with a low fringe. Overshoot is largest when the signal and fringe are gated on within a few milliseconds of each other, reaching 11.3 dB with the HF band and 3.8 dB with the LF band. These amounts of overshoot are close to the 13 and 3 dB reported by Schmidt and Zwicker (1991) under conditions with identical masker spectra.

With increasing onset asynchrony between the HF band and the signal, overshoot diminishes. When the signal is gated on after the HF onset, overshoot decreases to less than 1 dB for a 502-ms delay, the same amount of overshoot as with a continuous HF band. When the HF band is gated on after the signal, overshoot also diminishes with increasing onset asynchrony but no faster than when the HF band onset precedes signal onset. Even when the fringe comes on 18 ms after the signal, overshoot is still as large as 5 dB. The finding of pronounced overshoot for negative signal onset delays was unexpected; we call this overshoot at negative delays, "backward overshoot."

The amount of overshoot from a LF masker depends on signal delay in somewhat similar fashion to that from a HF masker, but the amount is much less, not exceeding on average a few decibels. One noteworthy difference between the LF and HF maskers is apparent for onset delays ranging from 22 to 502 ms, as well as with a continuous fringe. As seen in Fig. 2, threshold in these conditions is lower with the added LF band than without. This small negative overshoot becomes positive when the signal onset delay is 12 ms or less and remains approximately constant at 3 dB within a range of delays between –98 and 2 ms.

To examine the statistical significance of the data, a two-way ANOVA (masker spectral location × onset delay) for repeated measures was performed separately for parts 1, 2, and 3 of the experiment. The effect of fringe spectral location is significant in all three parts [part 1: $F(1,6)=1543.0$, $p \leqslant 0.0001$; part 2: $F(1,6)=2046.6$, $p \leqslant 0.0001$; part 3: $F(1,6)=2353.5$, $p \leqslant 0.0001$], as is the effect of the fringe delay [part 1: $F(6,36)=8.5$, $p \leqslant 0.0001$; part 2: $F(5,30)=53.0$, $p \leqslant 0.0001$; part 3: $F(7,42)=10.7$, $p \leqslant 0.0001$]. For the larger ranges of delay in parts 2 and 3, there also is a significant interaction between fringe spectral location and fringe delay [part 2: $F(5,30)=6.59$, $p<0.0005$; part 3: $F(7,42)=21.35$, $p \leqslant 0.0001$]. The effect of repetition is nonsignificant [part 1: $F(4,24)=0.72$, $p=0.59$; part 2: $F(4,24)=1.04$, $p=0.41$; part 3: $F(4,24)=1.76$, $p=0.17$].

IV. DISCUSSION AND CONCLUSIONS

The present results confirm the existence of a strong spectral asymmetry in overshoot: a low-frequency masker produces markedly less overshoot than a high-frequency masker. Moreover, overshoot from a high-frequency masker declines with signal delay in a manner consistent with the findings of Bacon and Viemeister (1985) for tone-on-tone masking. The present data, obtained with a noise masker, extend the findings of Bacon and Viemeister (1985) to the more typical overshoot paradigm based on noise maskers.

What is strikingly new is our demonstration that strong overshoot is observed when a high fringe band is added to an on-frequency masker after the signal offset. The backward overshoot found in the present study is reminiscent of offset overshoot, the threshold elevation observed when a brief signal is presented just before masker offset (e.g., Elliott, 1965; Elliott, 1967; Bacon and Viemeister, 1985).

The effects of backward overshoot and offset overshoot are similar as to their dependence on the masker spectrum. Experiments carried out with the use of tonal maskers (Bacon and Viemeister, 1985; Bacon and Moore, 1986) have demonstrated that offset overshoot is much stronger when the masker frequency lies above the signal frequency than when it is located below it. Another similarity between backward and offset overshoot is in the temporal course of these effects. Offset overshoot begins about 100 ms prior to the masker termination and increases as the signal is shifted toward the masker's end (Bacon and Viemeister, 1985). The

present data, obtained with a HF band, demonstrate that the temporal extent of backward overshoot is similar to that reported for offset overshoot.

The similarities in the spectral and temporal characteristics of backward and offset overshoot suggest that both these effects may be produced by similar processes. One possible explanation of these phenomena is the transient-masking effect (Bacon and Moore, 1987). Such an effect occurs when the signal is close to the onset or to the offset of a masker and the transient responses to the masker reduce the usefulness of transient responses to the signal.

Elliot (1965) ascribed the threshold elevation before masker offset to backward masking produced by central auditory offset responses. However, the offset and backward overshoot effects are unlikely to result from backward masking, because they depend on just the opposite frequency relation between signal and masker. Whereas backward masking is greater for a signal at a frequency above the masker's frequency band than for one below (Elliott, 1967), offset overshoot (Bacon and Viemeister, 1985) and backward overshoot are much greater when the signal frequency lies below the masker's frequency than when it lies above it.

Various hypotheses have been proposed in published studies to explain the physiological mechanism of overshoot [see Strickland, 2001 for a review]. Our data make it possible to verify certain hypotheses, especially those that explain the spectral asymmetry of overshoot.

Bacon and Moore (1987) ascribed the spectral asymmetry of overshoot to the transient-masking effect discussed earlier, in particular to the difference in the ratio of signal-evoked excitation to the masker-evoked excitation, in auditory channels above and below the signal frequency. They speculated that transient masking is centrally mediated, a notion that receives some support from the current demonstration of backward overshoot.

Another hypothesis links overshoot to active processing in the cochlea [see Strickland (2001) for a review]. It has been suggested that the active processes underlying the overshoot effect are mediated by the medial olivocochlear bundle of the efferent system. A primary basis for this hypothesis is that the time interval of 100–200 ms needed for the efferent system to reach its maximum effectiveness (Warren and Liberman, 1989) is similar to the interval between masker and signal onsets needed for threshold to decrease to its steady-state level (e.g., Zwicker, 1965a). The present data obtained in conditions with signal onset following fringe onset are in agreement with the efferent hypothesis. As seen in Fig. 2, the time interval after fringe onset during which threshold decreases to its steady-state level is about 200 ms.

The present findings have little bearing on the hypothesis that overshoot is caused by adaptation of auditory-nerve fibers (Smith and Zwislocki, 1975). Although most authors agree that neural adaptation contributes to the overshoot effect, it cannot account for the absence or near absence of overshoot when the masker does not include components in frequency regions above the signal frequency.

Another finding of the present experiment is worth noting. Thresholds were 2- to 3-dB higher with the CB masker alone (0-dB line in Fig. 2) than with the CB masker combined with the LF masker (circles) presented 500 to 20 ms earlier or continuously. A similar effect was also found by Schmidt and Zwicker (1991) with a continuous masker with a spectrum corresponding to the sum of our CB and LF maskers. A possible explanation for this threshold reduction is that neural adaptation caused by a preceding LF masker eliminated a small threshold elevation near the onset of the CB masker.

The inference of some overshoot with a narrowband noise is contrary to early investigations that suggested that masker spectral components located within the auditory channel centered on the signal frequency do not produce overshoot (Zwicker, 1965b). However, more recent experiments (Carlyon and White, 1992) demonstrate that this rule does not always hold.

ACKNOWLEDGMENTS

Comments by Dr. Sid P. Bacon, an anonymous reviewer, and the Associate Editor, Dr. Neal F. Viemeister, led to many improvements of the paper. The authors wish to express special thanks to Professor Bertram Scharf for his several stimulating discussions and suggestions that helped them greatly to shape the final version of the manuscript. Thanks also are due to Dr. Michael Epstein and Dr. Jeremy Marozeau for critically reading the manuscript. Søren Buus contributed substantially to this work, unfortunately he passed away prior to the submission of the revision. This work was supported by NIH/NIDCD Grant No. R01 DC 00187.

ANSI (**1989**). ANSI S3.6-1989, "Specifications for audiometers" (American National Standards Institute, New York).

Bacon, S. P., and Moore, B. C. J. (**1986**). "Temporal effects in simultaneous pure-tone masking: Effects of signal frequency, masker/signal frequency ratio, and masker level," Hear. Res. **23**, 257–266.

Bacon, S. P., and Moore, B. C. J. (**1987**). "Transient masking and the temporal course of simultaneous tone-on-tone masking," J. Acoust. Soc. Am. **81**, 1073–1077.

Bacon, S. P., and Smith, M. A. (**1991**). "Spectral, intensive, and temporal factors influencing overshoot," Q. J. Exp. Psychol. A **43A**, 373–400.

Bacon, S. P., and Viemeister, N. F. (**1985**). "The temporal course of simultaneous tone-on-tone masking," J. Acoust. Soc. Am. **78**, 1231–1235.

Carlyon, R. P., and White, L. J. (**1992**). "Effect of signal frequency and masker level on the frequency regions responsible for the overshoot effect," J. Acoust. Soc. Am. **91**, 1034–1041.

Elliott, L. L. (**1965**). "Changes in the simultaneous masked threshold of brief tones," J. Acoust. Soc. Am. **38**, 738–746.

Elliott, L. L. (**1967**). "Development of auditory narrow-band frequency contours," J. Acoust. Soc. Am. **42**, 143–153.

Hicks, M. L., and Bacon, S. P. (**1992**). "Factors influencing temporal effects with notched-noise maskers," Hear. Res. **64**, 123–132.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Schmidt, S., and Zwicker, E. (**1991**). "The effect of masker spectral asymmetry on overshoot in simultaneous masking," J. Acoust. Soc. Am. **89**, 1324–1330.

Smith, R. L., and Zwislocki, J. J. (**1975**). "Short-term adaptation and incremental responses of single auditory-nerve fibers," Biol. Cybern. **17**, 169–182.

Strickland, E. (**2001**). "The relationship between frequency selectivity and overshoot," J. Acoust. Soc. Am. **109**, 2062–2073.

Warren III, E. H., and Liberman, M. C. (**1989**). "Effects of contralateral sound on auditory-nerve responses. I. Contributions of cochlear efferents," Hear. Res. **37**, 89–104.

Zwicker, E. (**1965a**). "Temporal effects in simultaneous masking by white-noise bursts," J. Acoust. Soc. Am. **37**, 653–663.

Zwicker, E. (**1965b**). "Temporal effects in simultaneous masking and loudness," J. Acoust. Soc. Am. **38**, 132–141.

The Stokes relations linking time reversal and the inverse filter

F. Vignon, J.-F. Aubry, A. Saez, M. Tanter, D. Cassereau, G. Montaldo, and M. Fink
Laboratoire Ondes et Acoustique, ESPCI, Université Paris VII, U.M.R. C.N.R.S. 7587, 10 rue Vauquelin, 75005 Paris, France

(Received 10 March 2005; revised 5 November 2005; accepted 25 November 2005)

Our aim in this paper is to show how two adaptive focusing techniques, Time Reversal (TR) and the Spatio Temporal Inverse Filter (STIF), are related by the Stokes equations linking waves transmitted and reflected through a medium. For that purpose a model experiment has been investigated: a solid plate located between two arrays of transducers. When sending a wave from an array to the other through the plate, multiple transmitted waves are induced. TR and STIF are used to cancel these echoes. The echoes can be suppressed by TR, using the two arrays cavity surrounding the plate. They can also be cancelled by STIF, inverting the transmission operator between the arrays. The STIF achieves echoes cancellation by using only the transmitted fields through the plate, whereas TR also requires the reflected fields. The STIF's strategy is analyzed in light of the Stokes relations: thanks to the reflections in the medium, it is able to simulate a TR cavity with only one array. A mathematical analysis of the matrix expression of the Stokes relations then leads to two iterative ways to invert the transmission operator. Finally, this general technique is applied to a more complex medium: a human skull bone.. [DOI: 10.1121/1.2161452]

PACS number(s): 43.20.Fn, 43.20.El, 43.60.Tj [DRD] Pages: 1335–1346

I. INTRODUCTION

Time-reversal focusing techniques are based on the reversibility of acoustic propagation through complex media.[1–3] Acoustic reversibility means that for every burst of sound emitted from a source and possibly reflected and refracted by multiple interfaces or boundaries, there exists a set of waves that precisely retraces all the paths and converges in synchrony at the original source, as if time were going backward. This set of waves can be built using a time-reversal cavity: the medium under investigation is surrounded by elementary transducers located on a closed surface. Those transducers receive the signals originating from the source. Those signals, that can be long and complex due to multiple reflections/refractions in the medium, are then time reversed and sent back by the transducers; what has been received first is sent last. They converge toward the original location of the source, recreating at its location the pressure field that was first emitted. This idea leads to the concept of pulse compression through dispersive or reverberating media.[3,4]

Experimentally, it is possible to materialize a time-reversal cavity using two parallel linear ultrasound arrays[5] surrounding the medium of interest. In this paper, the medium of interest is a solid plate (Fig. 1): when emitting a wave from array 1, a series of echoes is recorded on array 2 due to multiple reflections at the solid–water interfaces.

On one hand, it has already been demonstrated that pulse compression and echoes cancellation can be achieved in closed time-reversal cavities by using a time-reversal operation.[1] On the other hand, it has been shown that if the cavity is not surrounding entirely the medium, a time-reversal operation does not cancel plate echoes whereas an inverse filter technique[6–9] achieves echo cancellation.[10] Our aim in this paper is to understand how one can use both arrays (a whole time-reversal cavity) or only one of them (a time-reversal mirror) in order to obtain pulse compression at a given location.

For that purpose, both methods are used: Time Reversal (TR) and the Spatio Temporal Inverse Filter (STIF). Time-reversal methods for pulse compression are presented in Sec. II. First time reversal is performed with a single Time-Reversal Mirror (TRM) in transmission or in reflection, and, second, with the whole Time-Reversal Cavity (TRC) made of both mirrors. Section III begins with a brief reintroduction of the STIF technique, then pulse compression results obtained by STIF with the TRM in transmission through the plate are presented. In Sec. IV, after a demonstration of the Stokes relations through a *gedanken* time-reversal experiment, it is shown how time reversal and the inverse filter are linked by these relations in a matrix formalism. This allows us to gain more understanding in both focusing methods, thus complementing the work done by Tanter *et al.*[11] about the comparison of TR and spatial inverse filter in terms of spatial focusing quality. Furthermore, it allows us to explain how the STIF, although using a TRM in transmission only, uses the Stokes relations to recover the information carried by the reflected waves otherwise lost by using time reversal with an unclosed cavity.

In Sec. V we show how the matrix expression of the Stokes relations leads to an iterative approach to reach inverse filter focusing, giving a new theoretical basis to the intuitive iterative time-reversal technique introduced recently by Montaldo *et al.*[12] Finally in Sec. VI the general considerations made in all previous sections are applied to a more complex problem: spatial focusing through a human skull. The experimental results demonstrate that the theory presented in this paper is very general and can be applied to objects more complicated than the one-dimensional plate used here to illustrate the theory.

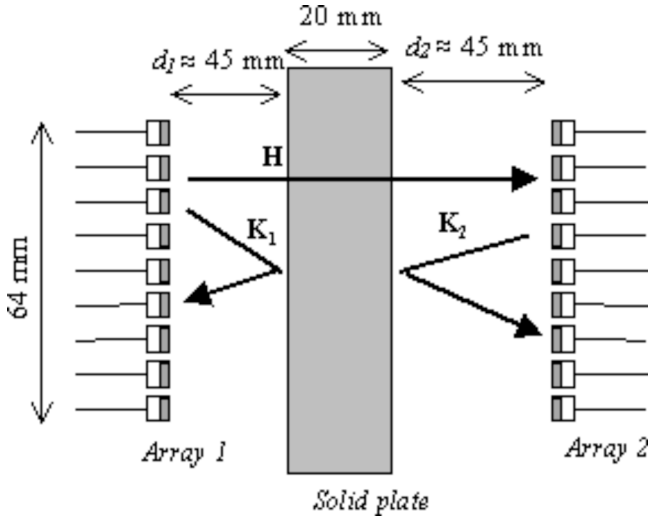


FIG. 1. Description of the model experiment. Two linear arrays are placed parallel to the faces of a titanium plate. The two reflection propagators, from array 1 to itself $[\mathbf{K}_1(\omega)]$, from array 2 to itself $[\mathbf{K}_2(\omega)]$, and the transmission propagator from array 1 to array 2 $[\mathbf{H}(\omega)]$, are acquired.

II. THE TIME-REVERSAL MIRROR (TRM) AND THE TIME-REVERSAL CAVITY (TRC)

A. Propagation operators

The experimental setup is presented in Fig. 1: a solid titanium plate (20 mm width, longitudinal sound speed 5.765 mm μs^{-1}) is immersed in water and surrounded by two identical linear arrays parallel to it. Each array is made of 64 identical transducers working at 1.5 MHz central frequency (60% bandwidth), with a 1 mm spatial pitch between two adjacent transducers, so that the lateral dimension of the arrays is 64 mm. The lateral dimension of the plate is much greater than the lateral dimensions of the arrays so that the plate can be considered to be of infinite lateral dimensions. The transducers measure 1.2 cm (12 times the central wavelength in water) in the direction perpendicular to the plane illustrated in Fig. 1. This allows us to consider wave propagation only in the plane containing the arrays. The arrays are located, respectively, at a distance $d_1 \approx d_2 \approx 45$ mm from the plate.

The transmission of waves from one array to the other and reflection from one array to itself can be formalized. For each pair {element m on array 2, element j on array 1}, the impulse response $h_{mj}(t)$ is defined as the signal received on the mth element of array 2 after a temporal Dirac function is emitted by the jth element of array 1. This response includes all the propagation effects through the medium as well as the acoustoelectric responses of both transducers. If each element j of array 1 emits a temporal signal $e_j(t)$ $(1 \leqslant j \leqslant \mathrm{N})$, the output signal $f_m(t)$, $1 \leqslant m \leqslant \mathrm{N}$ received by element m of array 2 is

$$f_m(t) = \sum_{j=1}^{N} h_{mj}(t) \otimes e_j(t), \quad 1 \leqslant m \leqslant \mathrm{N}, \tag{1}$$

where $\otimes$ is the temporal convolution operator.

A Fourier transform leads to the following relation:

$$F_m(\omega) = \sum_{j=1}^{N} H_{mj}(\omega) E_j(\omega), \quad 1 \leqslant m \leqslant \mathrm{N}, \tag{2}$$

which can be written in matrix formalism as

$$F(\omega) = \mathbf{H}(\omega) E(\omega). \tag{3}$$

The propagation operator $\mathbf{H}(\omega)$ is an $N \times N$ matrix that contains the impulse responses between any element j of array 1 and m of array 2. It describes the propagation in the medium from array 1 to array 2.

In a first step, the transmission propagators $\mathbf{H}(\omega)$ from array 1 to array 2, and the reflection propagators $\mathbf{K}_1(\omega)$ from array 1 to itself and $\mathbf{K}_2(\omega)$ from array 2 to itself (Fig. 1), are measured experimentally. Note that because of spatial reciprocity the reflection propagators are equal to their transpose $[{}^t\mathbf{K}_1(\omega) = \mathbf{K}_1(\omega), {}^t\mathbf{K}_2(\omega) = \mathbf{K}_2(\omega)]$.

B. The time-reversal mirror (TRM)

The aim is to obtain on array 2 a single-front plane wave with a given temporal shape $s(t)$ and an associated spectrum $s(\omega)$. The temporal shape of this "target" signal is presented in Fig. 2(a), and its spectrum in Fig. 2(b). Its spectrum matches the frequency response of the transducers. Array 1 is used as a TRM: the target signal $S_2(t) = s(t)^t[1, 1, \ldots, 1]$ is first emitted by array 2. According to the previous section, the signals recorded on array 1 are given by

$$R_1(\omega) = {}^t\mathbf{H}(\omega) S_2(\omega). \tag{4}$$

Indeed, as $\mathbf{H}(\omega)$ is the propagator from array 1 to 2, the propagator from array 2 to 1 is its transpose ${}^t\mathbf{H}(\omega)$. This is due to spatial reciprocity: $h_{mj}^{2\to1}(t) = h_{jm}^{1\to2}(t)$.

The temporal shape of the signals R_1 received on array 1 is given in Fig. 3(a): the main wave front is followed by a series of echoes resulting from the reflection of the wave front on the interfaces between the plate and the water. (Every time the temporal shape of signals received or emitted by an array is presented, only the signals corresponding to the central element of the array are shown: these are the cleanest signals, free from edge effects. However, they are representative of the signals received on all other elements since "quasi"-plane waves are used.) Those signals are then time reversed and sent back toward array 2. This time-reversal operation is mathematically described by the variable change $(t) \to (-t)$ that is equivalent, in the frequency domain, to a phase conjugation operation. The reemitted signals are then given by

$$E_1(\omega) = R_1(\omega)^* = {}^t\mathbf{H}(\omega)^* S_2(\omega)^*. \tag{5}$$

These signals $E_1(\omega)$ are propagated toward array 2, and the signals recorded on array 2 are finally given by

$$F_2(\omega) = \mathbf{H}(\omega) E_1(\omega) = [\mathbf{H}(\omega) {}^t\mathbf{H}(\omega)^*] S_2(\omega)^*. \tag{6}$$

The temporal shape of the signals $f_2(t)$ received on array 2 at the end of this time-reversal process is presented in Fig. 3(b). Because of multiple reflections of the wave in the plate, the signals present a series of echoes preceding and following the main wave front. On the spectrum of those signals [Fig. 3(c), solid curve], an enhancement of the frequencies

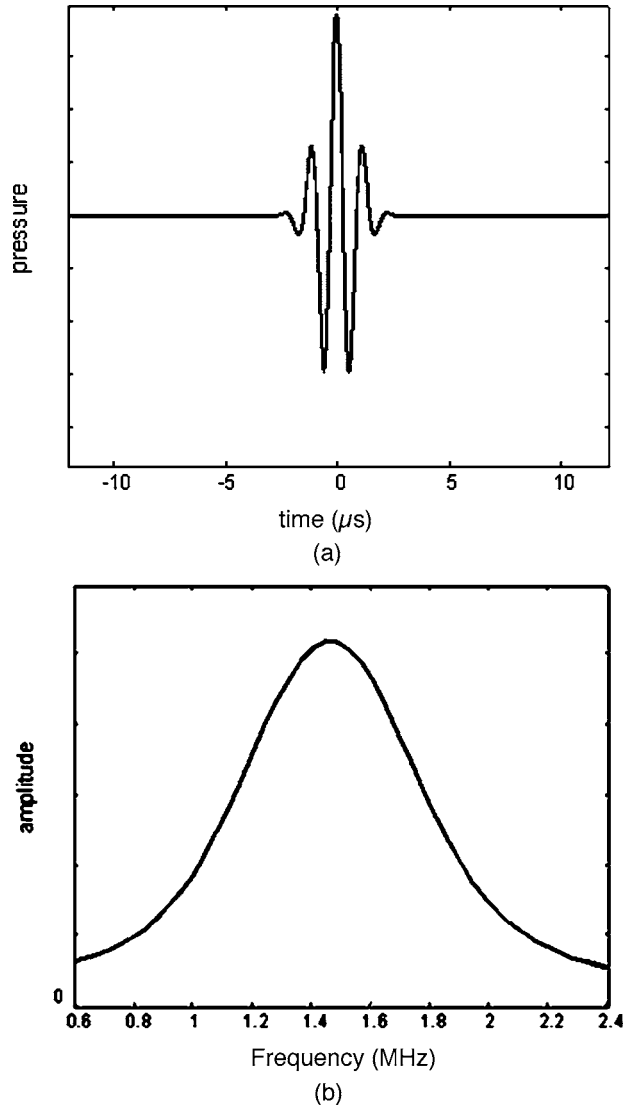


FIG. 2. Temporal shape $s(t)$ (a) and spectrum $s(\omega)$ (b) of the "target" signals that have to be received simultaneously on all elements of array 2. The spectrum of the target signal has been chosen to be close to the frequency response of the transducers used in the experiments.

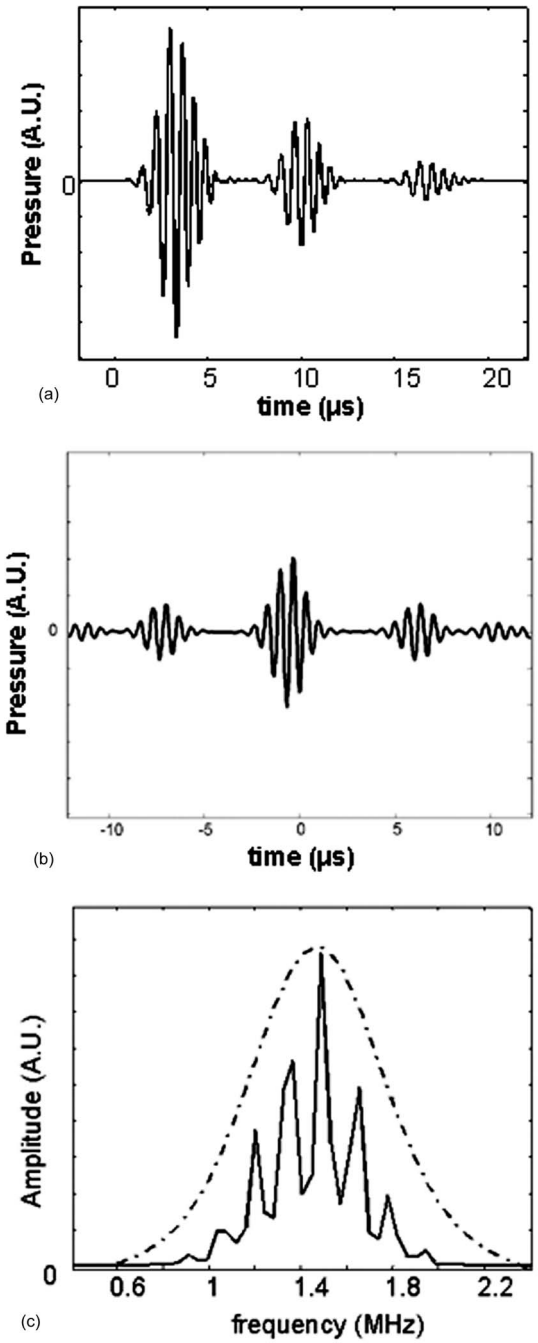


FIG. 3. (a) The temporal signal received on the central transducer of array 1 after the target signal is sent from array 2. (b) The temporal signal received on the central transducer of array 2 after time reversal in transmission (using only array 1 to record, time reverse, and send back the waves). (c) Normalized spectrum of the signal received on array 2 after time reversal in transmission is performed (solid curve) along with the normalized spectrum of the target signals (dash–dotted curve).

that are resonant in the plate is observed: it is far from the spectrum $S_2(\omega)$ of the target signal (dash-dotted curve).

A time-reversal operation performed with a time-reversal mirror cannot achieve pulse compression: this is due to the limited extension of the TRM, confirming the results obtained in Ref. 10. The reflected wave front going back to array 2 is not recorded by array 1 and the information that it contains is lost for the TR process. The reflected wave also needs to be recorded and time reversed in order to obtain a good pulse compression. This can be achieved by using both arrays as a closed time-reversal cavity.

C. The time-reversal cavity (TRC)

The experimental setup allows us to extend the time-reversal mirror to a nearly closed time-reversal cavity by using both arrays in the time-reversal process. The target signal $S_2(\omega)$ is emitted by array 2 and the corresponding signals $R_1(\omega)$ and $R_2(\omega)$ are recorded on both arrays 1 and 2, time reversed, and sent back simultaneously. The resulting signal $F_2(\omega)$ then received on array 2 is given by

$$F_2(\omega) = [\mathbf{H}(\omega)^{\mathrm{t}}\mathbf{H}(\omega)^*]S_2(\omega)^* + [\mathbf{K}_2(\omega)^{\mathrm{t}}\mathbf{K}_2(\omega)^*]S_2(\omega)^*. \tag{7}$$

The first half of the above expression describes the TRM process in transmission using only array 1, and the second half in reflection using only array 2.

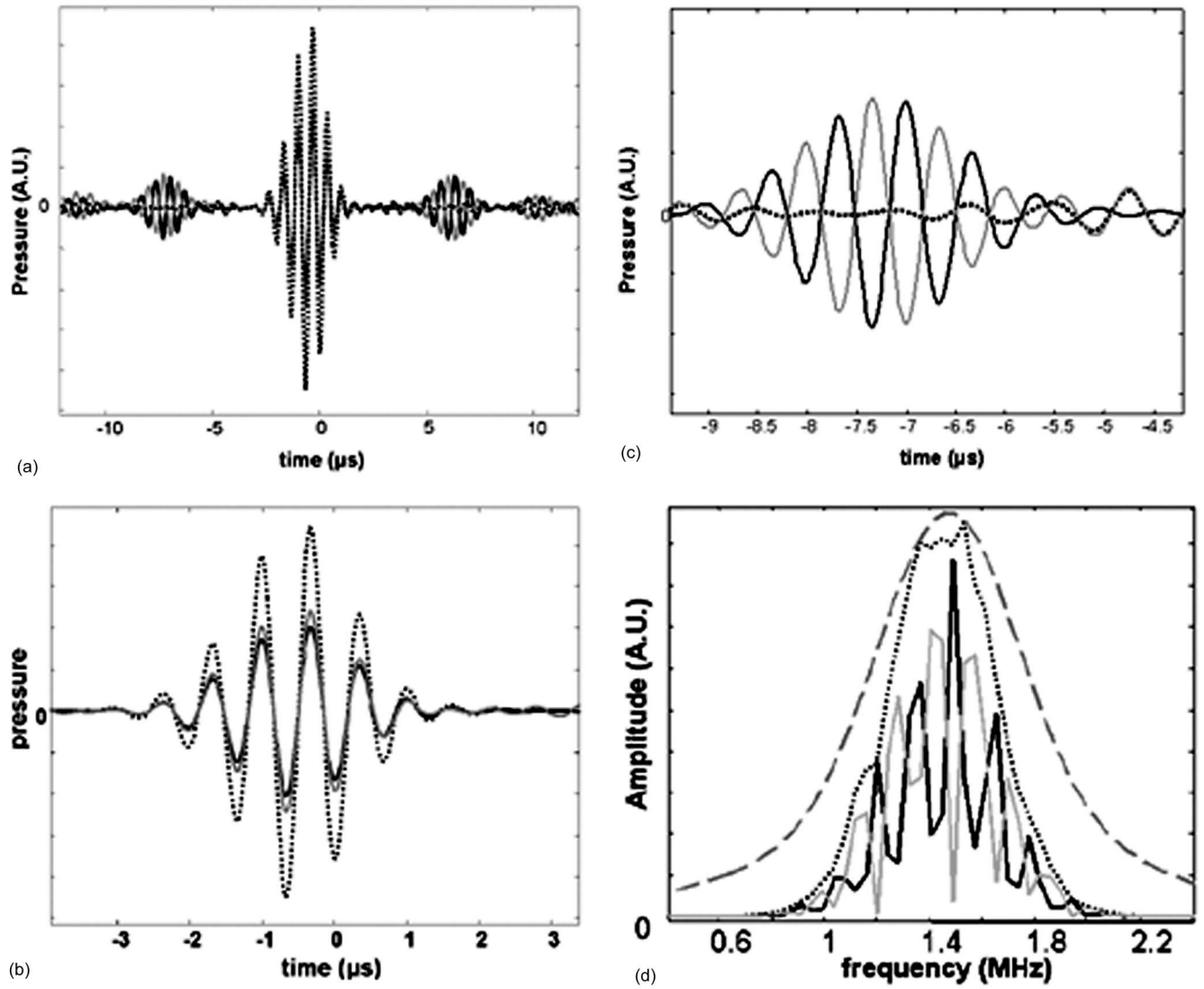


FIG. 4. Temporal shape $f_2(t)$ (a)–(c) and spectrum $F_2(\omega)$ (d) of the signals received on the central transducer of array 2 after time-reversal processes. Dotted curve: using the whole time-reversal cavity (both arrays 1 and 2). Black curve: using the time-reversal mirror in transmission (only array 1). Gray curve: TRM in reflection (using only array 2). (a) presents a long scan spanning the main wave front and two echoes preceding and following the main front. (b) is a temporal zoom on the main wave front where transmission and reflection TRM signals interfere constructively and (c) the first echo preceding the main wave front, where transmission and reflection TRM signals interfere destructively. (d) presents the corresponding spectra: the dotted spectrum is the signal $F_2(\omega)$ recorded on array 2 when the whole TRC is used, the black spectrum when only the TRM is transmission is used, and the gray spectrum when only the TRM is reflection is used. The gray dashed curve is the spectrum of the "target" signal $s(\omega)$.

Three experiments have been conducted. First, array 1 only is emitting the time-reversed signals. The corresponding signals received on array 2 are plotted in Fig. 4(a) (black line), they correspond to the results obtained in the previous section with the TRM in transmission. Second, array 2 only is emitting the time-reversed signals (TRM in reflection). The corresponding signals received on array 2 are superimposed in Fig. 4(a) (gray line): those signals appear to be in phase with the signals of the TRM in transmission for the main wave front at time $t=0$, whereas they are in phase opposition for the echoes. In a last experiment, both arrays emit simultaneously their time-reversed signal (dotted line) confirming the expected echo cancellation when the whole TRC is used. Both signals originating from the TRM in transmission and in reflection interfere constructively at the arrival time of the main wave front [Figs. 4(a) and 4(b)], and destructively at the arrival times of the echoes [Figs. 4(a) and 4(c)], thus leading to main front enhancement and echoes cancellation. In the frequency domain [Fig. 4(d)] one can see that frequencies that are enhanced by the TRM in transmission (black line) are the ones that are lost by the TRM in reflection (gray line). The spectrum of the signals obtained on array 2 when using the whole TRC (dotted line) correspond to the sum of the spectra obtained with both TRMs, it approaches the target signal spectrum (gray dashed line).

III. THE SPATIOTEMPORAL INVERSE FILTER (STIF)

In this section we show how a pulse compression as good as the one obtained with the whole TRC can be obtained using only one emitting array (array 1).

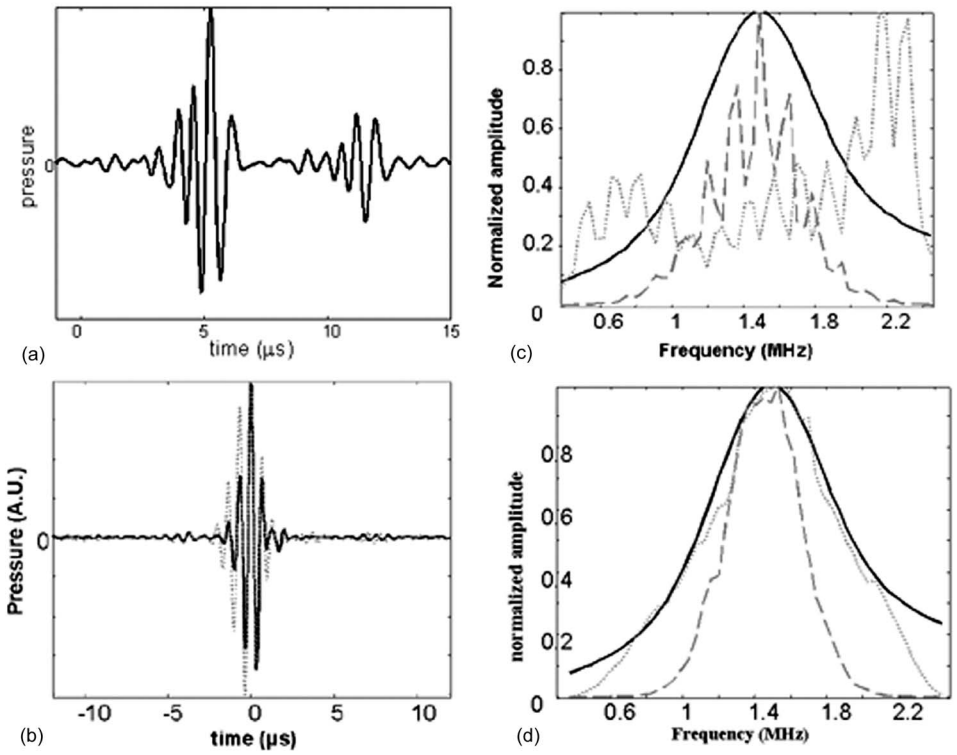


FIG. 5. (a) The signal $e_1(t)$, calculated by the inverse filter, to be emitted from array 1 in order to focus on the target signal $s(t)$ on array 2. (b) The resulting signals $f_2(t)$ obtained on the central element of array 2 (black curve), along with the signals focused by using the TRC (gray dotted curve). (c) The inverse filter's emission signals' spectrum $E_1(\omega)$ (gray dotted curve), to be compared to the transmission TRM emission signals (gray dashed cure) and target signals (black curve). (d) The spectrum of the focused signals by IF (gray, dotted) along with the spectra of the focused signals with TRC (gray, dashed) and the target signal (black).

As introduced in Sec. II, when a set of signals $E_1(\omega)$ is emitted by array 1, the set of signals $F_2(\omega)$ received on array 2 after propagation through the plate is linked to the emitted signals by the propagator $\mathbf{H}(\omega)$:

$$F_2(\omega) = \mathbf{H}(\omega)E_1(\omega). \tag{3'}$$

The aim is to determine the desired set of signals $E_1(\omega)$ that would give rise after propagation to the target signal $S_2(\omega)$ on array 2.

For this purpose, Eq. (3) can be directly inverted, and the signals to emit are

$$E_1(\omega) = \hat{\mathbf{H}}(\omega)^{-1}S_2(\omega). \tag{8}$$

This approach is called the Spatio Temporal Inverse Filter (STIF) technique.[6–9] As is usually the case in inverse problems, the inversion of $\mathbf{H}(\omega)$ is ill conditioned: $\hat{\mathbf{H}}(\omega)^{-1}$ is a regularized inverse called pseudoinverse, obtained inverting only the physically significant eigenvalues of $\mathbf{H}(\omega)$. The singular value decomposition of $\mathbf{H}(\omega)$ is performed at each frequency within the bandwidth of the transducers, only the highest eigenvalues are inverted, the lowest eigenvalues (below a given threshold chosen by the user) are not inversed but set to 0.[9]

The final signals $e_1(t)$ to be emitted from array 1 are presented in Fig. 5(a). They are obtained after going back to the time domain by an inverse Fourier transform. By analyzing the temporal shape of these signals, one can understand the STIF's strategy to eliminate the echoes: first, a main wave front is sent. This wave front will produce the desired signal on array 2. It will also give rise to a reflected signal at the second solid–water interface. A second wave front of smaller amplitude is then sent in order to interfere destructively with this reflected wave front, thus eliminating the echoes that would otherwise follow the main wave front.

Looking at the spectrum of the emitted signals $E_1(\omega)$ [Fig. 5(c), gray dotted curve] allows us to give another interpretation of the STIF's strategy: the frequencies that will be enhanced by the presence of the plate (the resonant frequencies) are sent with a low amplitude, whereas the frequencies that will be attenuated—because of the limited bandwidth of the transducers and by passing through the plate—are sent with a stronger amplitude: the STIF achieves an amplitude compensation in the frequency domain.

The STIF signals $E_1(\omega)=\hat{\mathbf{H}}(\omega)^{-1}S_2(\omega)$ are sent from array 1 and the signals $F_2(\omega)$ subsequently received on array 2 are

$$F_2(\omega) = \mathbf{H}(\omega)E_1(\omega) = [\mathbf{H}(\omega)\hat{\mathbf{H}}(\omega)^{-1}]S_2(\omega). \tag{9}$$

The received signals $F_2(\omega)$ would be equal to the target signals $S_2(\omega)$ if the inversion was perfect. The experimental temporal signals in focal plane $f_2(t)$ are indeed close to the target signals $s(t)$, as shown in Fig. 5(b) (plain curve). In particular, echo cancellation is achieved in spite of the fact that only one array is used: the STIF achieves with a single array a pulse compression equivalent to the one obtained with the two arrays of the TRC. For comparison, in Fig. 5(b) the focused signals with echo cancellation obtained with the TRC (gray dotted curve) have also been plotted. Figure 5(d) presents the spectra of the signals received in focal plane when trying to obtain the spectrum of the target signals $S_2(\omega)$ (black curve) when using the TRC (gray dashed curve) and the STIF (gray dotted curve). The STIF focusing is even closer to the target than the time-reversal cavity's focusing. This is due to the fact that the STIF enhances the frequencies

that are not well transmitted by the transducers [see Fig. 5(c)], thus achieving a bandwidth compensation. The effects of this bandwidth compensation are visible when comparing the temporal length of the focused wave front by STIF or by the TRC [Fig. 5(b)]: the STIF focused signals are slightly shorter than the TRC focused signals.

IV. HOW DOES THE STIF MIMIC A COMPLETE TIME-REVERSAL CAVITY WITH ONE SINGLE ARRAY?

The STIF is only based on the transmission propagator **H**, but it is able to cancel the echoes with the same efficiency as the time-reversal cavity based on **H** and $\mathbf{K}_2$. How can the STIF retrieve in the transmitted field the information contained in the missing reflected field?

The link between the time-reversal cavity and the spatiotemporal inverse filter originates, in fact, in the Stokes relations.

A. The Stokes relations demonstrated by a *gedanken* time-reversal experiment

The implication of time-reversal invariance was studied for the first time by Stokes in the framework of the most elementary model of a complex medium: a plane interface separating two media with different wave velocities [Fig. 6, this setup has more recently been further theoretically investigated in Refs. 13 and 14 for spatial focusing purposes].

An incident plane wave with an amplitude normalized to 1 propagates toward the interface from medium 2 to medium 1. It gives rise to both a reflected plane wave of amplitude r_{22} and a transmitted wave of amplitude t_{12} [Fig. 6(a)]. The consequence of time-reversal symmetry is that the time-reversed experiment is also valid. It can be described by a new set of three waves: two incident waves of amplitude r_{22} and t_{12} propagating in medium 2 and medium 1, respectively, and a wave of amplitude 1 propagating in medium 2 away from the interface [Fig. 6(b)]. On the other hand, defining r_{11} and t_{21} as the reflection and transmission coefficients for an incident wave coming from medium 1, the superposition principle shows that these two incident waves lead to the generation of four plane waves propagating away from the interface, two of them propagating in medium 2 with a resulting amplitude $t_{21}t_{12}^{*}+r_{22}r_{22}^{*}$, and the two other propagating in medium 1 with a resulting amplitude $r_{11}t_{12}^{*}+t_{12}r_{22}^{*}$ [Fig. 6(c)]. As the amplitudes of these waves must be equal to the ones of the waves in the time-reversed experiment [the situations described in Figs. 6(b) and 6(c) must be equivalent], it allows us to verify the Stokes relations:

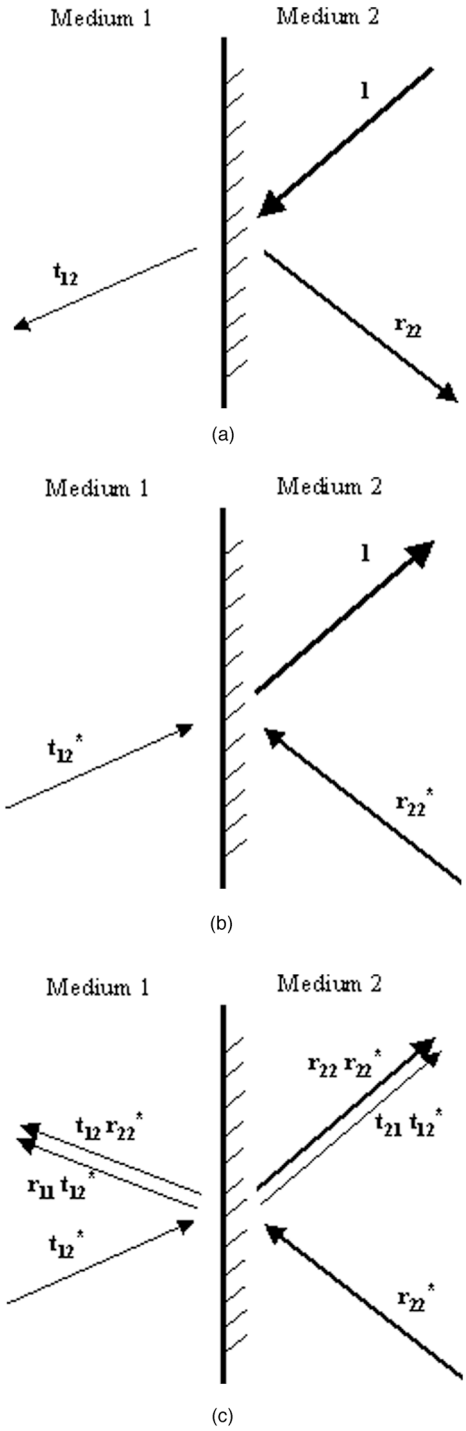


FIG. 6. Description of Stokes' *gedanken* time-reversal experiment. (a) An incident wave is sent toward an interface between two different media; it gives rise to a transmitted and a reflected wave. (b) the time-reversed experiment, highlighting its similarities with the direct experiment. (c) The time-reversed experiment, showing all the reflected and transmitted waves. The identification of the situations described in (b) and (c) leads to the Stokes relations.

$$r_{11}t_{12}^{*} + t_{12}r_{22}^{*} = 0 \qquad \text{(S1) (resulting from the identification of the waves propagating in medium 1);}$$

$$t_{21}t_{12}^{*} + r_{22}r_{22}^{*} = 1 \qquad \text{(S2) (resulting from the identification of the waves propagating in medium 2).}$$

It can be demonstrated in a similar way, or by analogy with (S1) and (S2) that also

$$r_{22}t_{21}^{*} + t_{21}r_{11}^{*} = 0 \qquad \text{(S3)},$$
$$t_{12}t_{21}^{*} + r_{11}r_{11}^{*} = 1 \qquad \text{(S4)}.$$

B. The Stokes relations in a matrix formalism

The same approach can be applied in a matrix formalism thanks to the propagation operators introduced in Sec. I. If a set of signals E_1 and E_2 are, respectively, emitted by arrays 1 and 2, the corresponding signals R_1 and R_2 recorded, respectively, by arrays 1 and 2 are given by

$$\begin{pmatrix} R_1 \\ R_2 \end{pmatrix} = \begin{pmatrix} \mathbf{K}_1 & {}^{t}\mathbf{H} \\ \mathbf{H} & \mathbf{K}_2 \end{pmatrix} \begin{pmatrix} E_1 \\ E_2 \end{pmatrix} = \mathbf{H}_{\text{global}} \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}. \tag{10}$$

$\mathbf{H}_{\text{global}}$ is the global propagation operator from arrays 1 and 2 to arrays 1 and 2. For the sake of simplicity, the frequency dependence has not been explicitly written. $\mathbf{H}_{\text{global}}$ corresponds to the so-called partitioned $\mathbf{S}$ matrix in optics.[15,16]

Let us assume, as in Sec. II, that the target signals S_2 are emitted by array 2. It gives rise to a reflected (R_2) and transmitted (R_1) wave front:

$$\begin{pmatrix} R_1 \\ R_2 \end{pmatrix} = \mathbf{H}_{\text{global}} \begin{pmatrix} 0 \\ S_2 \end{pmatrix}. \tag{11}$$

If those signals are time reversed and propagate again through $\mathbf{H}_{\text{global}}$, one will finally record, on the arrays,

$$\begin{aligned} \begin{pmatrix} R_{TR1} \\ R_{TR2} \end{pmatrix} &= \mathbf{H}_{\text{global}} \begin{pmatrix} R_1^{*} \\ R_2^{*} \end{pmatrix} = \mathbf{H}_{\text{global}}\mathbf{H}_{\text{global}}^{*} \begin{pmatrix} 0 \\ S_2^{*} \end{pmatrix} \Leftrightarrow \begin{pmatrix} R_{TR1} \\ R_{TR2} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{K}_1\mathbf{K}_1^{*} + {}^{t}\mathbf{H}\mathbf{H}^{*} & \mathbf{K}_1{}^{t}\mathbf{H}^{*} + {}^{t}\mathbf{H}\mathbf{K}_2^{*} \\ \mathbf{H}\mathbf{K}_1^{*} + \mathbf{K}_2\mathbf{H}^{*} & \mathbf{H}{}^{t}\mathbf{H}^{*} + \mathbf{K}_2\mathbf{K}_2^{*} \end{pmatrix} \begin{pmatrix} 0 \\ S_2^{*} \end{pmatrix} \\ &= \begin{pmatrix} (\mathbf{K}_1{}^{t}\mathbf{H}^{*} + {}^{t}\mathbf{H}\mathbf{K}_2^{*})S_2^{*} \\ (\mathbf{H}{}^{t}\mathbf{H}^{*} + \mathbf{K}_2\mathbf{K}_2^{*})S_2^{*} \end{pmatrix}. \end{aligned} \tag{12}$$

According to the time-reversal invariance, one should recover the time-reversed version of the initial field, so that $R_{TR1}=0$ and $R_{TR2}=S_2^{*}$. Consequently,

$$\mathbf{K}_1{}^{t}\mathbf{H}^{*} + {}^{t}\mathbf{H}\mathbf{K}_2^{*} = \mathbf{0} \qquad \text{(S1')},$$
$$\mathbf{H}{}^{t}\mathbf{H}^{*} + \mathbf{K}_2\mathbf{K}_2^{*} = \mathbf{I} \qquad \text{(S2')},$$

where $\mathbf{I}$ is the identity operator. The two previous equations are the Stokes relations in matrix formalism. Note the analogy of (S1') and (S2') with the scalar relations (S1) and (S2), the transmission operator $\mathbf{H}$ being analogous to the transmission coefficient t_{21}, its transpose ${}^{t}\mathbf{H}$ to t_{12}, and the reflection operators $\mathbf{K}_1$ and $\mathbf{K}_2$ to the reflection coefficients r_{11} and r_{22}, respectively.

It can be demonstrated in a similar way, or by analogy with (S1') and (S2'), that also

$$\mathbf{K}_2\mathbf{H}^{*} + \mathbf{H}\mathbf{K}_1^{*} = \mathbf{0} \qquad \text{(S3')},$$
$${}^{t}\mathbf{H}\mathbf{H}^{*} + \mathbf{K}_1\mathbf{K}_1^{*} = \mathbf{I} \qquad \text{(S4')}.$$

The same analogy can be done between (S3') and (S4') and their scalar counterparts (S3) and (S4). (S1'), (S2'), (S3'), and (S4') can be summed up in a single matrix relation:

$$\mathbf{H}_{\text{global}}\mathbf{H}_{\text{global}}^{*} = \mathbf{I}. \qquad \text{(S)}$$

This formalizes the fact that the two time-reversal mirrors make up a perfect time-reversal cavity (a time-reversal experiment performed with the two arrays leads exactly to the original signal: the global time-reversal operator is equal to identity).

C. How does the inverse filter technique mimic a complete time-reversal cavity with one single array?

Multiplying the second Stokes relation (S2') on the left by $\mathbf{H}^{-1}$, one finally obtains

$${}^{t}\mathbf{H}^{*} + \mathbf{H}^{-1}\mathbf{K}_2\mathbf{K}_2^{*} = \mathbf{H}^{-1}. \tag{13}$$

This relation links the Inverse Filter with the Time-Reversal Cavity. Indeed, the set of signals E_{IF}, calculated with the inverse filter, to be emitted by array 1 in order to focus on target S_2 on array 2 are

$$\begin{aligned} &E_{\text{IF}} = \mathbf{H}^{-1}S_2 \\ &\Leftrightarrow E_{\text{IF}} = {}^{t}\mathbf{H}^{*}S_2 + \mathbf{H}^{-1}\mathbf{K}_2\mathbf{K}_2^{*}S_2 \\ &\Leftrightarrow E_{\text{IF}} = E_{\text{TR}} + E', \end{aligned} \tag{14}$$

$E_{\text{TR}} = {}^{t}\mathbf{H}^{*}S_2$ correspond to the signals that are emitted by array 1 in order to focus on S with the TRM in transmission. $E' = \mathbf{H}^{-1}\mathbf{K}_2\mathbf{K}_2^{*}S_2$ is harder to understand. $\mathbf{K}_2\mathbf{K}_2^{*}S_2$ corresponds to the signals that would have been emitted by array 2 if the TRM in reflection was used to focus properly on S_2. Those signals are multiplied by $\mathbf{H}^{-1}$ to adapt these signals in order to emit them from array 1 instead of array 2. In other words, E_{IF} contains all the information recorded by the two arrays time-reversal cavity but can be emitted with one single array (array 1). In fact, thanks to the Stokes relations, transmitted and reflected wave fronts are directly linked, giving the opportunity to recover information on the reflected wave front from the transmitted wave front. In the model experiment presented here, one can understand that this information is in one sense stored in time as the multiple echoes are the signature of the presence of the solid–water interfaces.

This explains *a posteriori* why the STIF is able to cancel the echoes by using only one emitting array.

V. ITERATIVE INVERSION

A. Simple iteration method

Equation (13) gives $\mathbf{H}^{-1}$ as a function of $\mathbf{H}^{-1}$. It suggests the idea of using a matrix sequence $\{\mathbf{M}_p\}$ to compute $\mathbf{H}^{-1}$. The matrix sequence is defined as

$$\begin{aligned} &\mathbf{M_0} = \mathbf{0}, \\ &\mathbf{M}_{p+1} = {}^{t}\mathbf{H}^{*} + \mathbf{M}_p\mathbf{K}_2\mathbf{K}_2^{*}. \end{aligned} \tag{15}$$

It can be demonstrated mathematically (see Appendix A) that this sequence converges to $\mathbf{H}^{-1}$. This sequence is an iterative way to invert any propagator $\mathbf{H}$. Note that the propagator $\mathbf{K}_2\mathbf{K}_2^{*}$ that intervenes in the definition of the sequence does not really need to be measured: indeed, the operator $\mathbf{K}_2\mathbf{K}_2^{*}$ is known if $\mathbf{H}$ is known, via the Stokes relation (S2'): $\mathbf{K}_2\mathbf{K}_2^{*}$

$= \mathbf{I} - \mathbf{H}^{t}\mathbf{H}^{*}$, and the sequence (15) can be redefined with no need of $\mathbf{K}_2$:

$$\mathbf{M}_0 = \mathbf{0},$$

$$\mathbf{M}_{p+1} = {}^{t}\mathbf{H}^{*} + \mathbf{M}_p(\mathbf{I} - \mathbf{H}{}^{t}\mathbf{H}^{*}). \quad (15')$$

In order to check the convergence of sequence (15), $\mathbf{H}$ and $\mathbf{K}_2$ have been experimentally measured and each term $\mathbf{M}_p$ of the sequence have been successively computed thanks to (15), at all frequencies in the bandwidth of the transducers. The convergence can be visualized by looking at the eigenvalues of the product $\mathbf{HM}_p$ at a given frequency for different iteration numbers. The product $\mathbf{HM}_p$ is expected to converge to the identity matrix, so its eigenvalues are expected to converge to 1.

These eigenvalues are plotted in Fig. 7(a) at the central frequency for 2, 4, 6, 8, 10 iterations. Each curve has been labeled with the number of iterations. Only the 15 highest eigenvalues converge to 1, whereas the noise is kept close to zero. They correspond to physically significant eigenvalues emerging from noise. It has been shown in Ref. 9 that the number of eigenvalues corresponds to the number of independent focal spots that the array is able to generate in the aperture of the image plane. When performing TR in transmission to focus from array 1 on the central element of array 2, the corresponding amplitude distribution recorded on array 2 is plotted in Fig. 7(b). The main lobe lateral width is 4 mm. As the array is 64 mm wide, this means that it is possible to obtain 64 mm/4 mm=16 independent focal spots on array 2 with this experimental setup. Due to edge effects, one would expect the rank of the propagator to be equal to 16 or slightly below. This is in good agreement with the observed 15 nonzero eigenvalues.

Along with the eigenvalues of the product $\mathbf{HM}_p$ are plotted the eigenvalues of the product $\mathbf{H}\hat{\mathbf{H}}^{-1}$ at the same frequency: in the regularized inversion, only the 15 first eigenvalues have been inverted because the remaining eigenvalues are too weak: the rank of matrix $\mathbf{H}$ is 15 at this frequency. This explains why the eigenvalues of $\mathbf{H}\hat{\mathbf{H}}^{-1}$ look like a Heaviside function with a cutoff at the 15th eigenvalue.

Another way to check the convergence is to apply the operator $\mathbf{HM}_p$ to the target signal S_2, for different values of p. [This operation (that is repeated in this section and in Sec. V) is performed in the computer only: the vector $\mathbf{H}^{t}\mathbf{H}^{*}S_2$ is calculated, and then its Fourier transform is presented.] For $p=0$, the vector $\mathbf{HM}_0S_2$ is equal to $\mathbf{H}^{t}\mathbf{H}^{*}S_2$, i.e., the focusing obtained by time reversal in transmission. The limit of the sequence $\mathbf{HM}_pS_2$ should be equal to $\mathbf{HH}^{-1}S_2$, i.e., the focusing by the spatiotemporal inverse filter. The temporal shapes of the vectors $\mathbf{HM}_pS$ for $p=0,3,10$ are presented in Fig. 7(c). When iterating, the echoes preceding and following the main wave front disappear. Furthermore, the main wave front gets shorter, evidencing the bandwidth compensation performed by the STIF.

It is interesting to note that this iterative inversion is formally equivalent to the iterative time reversal introduced recently by Montaldo *et al.*[12] Their intuitive approach consisted in propagating experimentally the difference between the desired focusing and the time-reversal focusing.

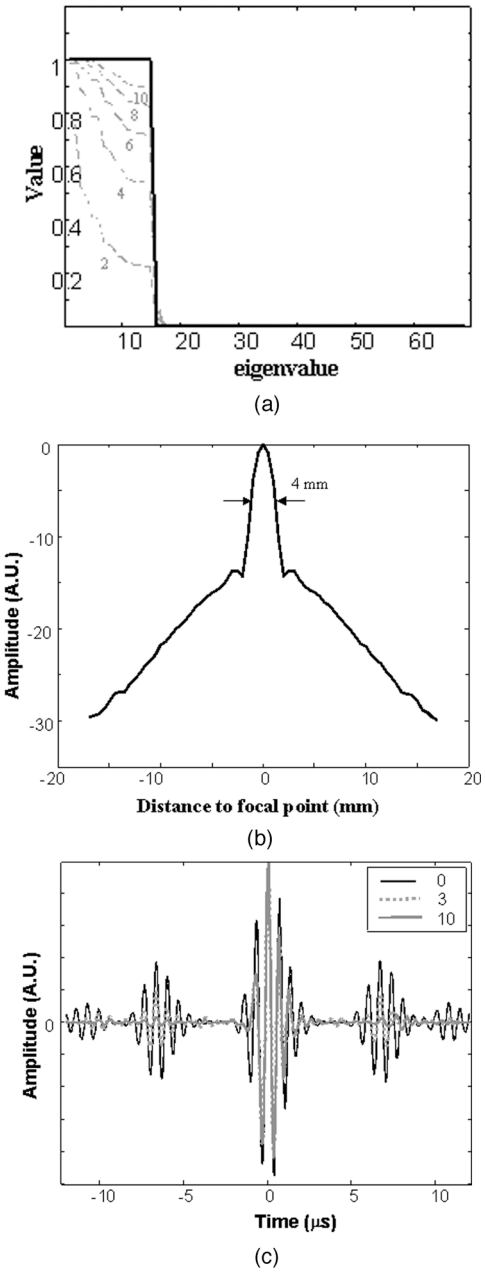
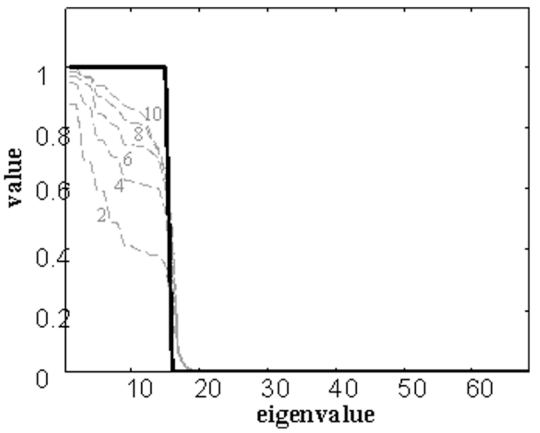


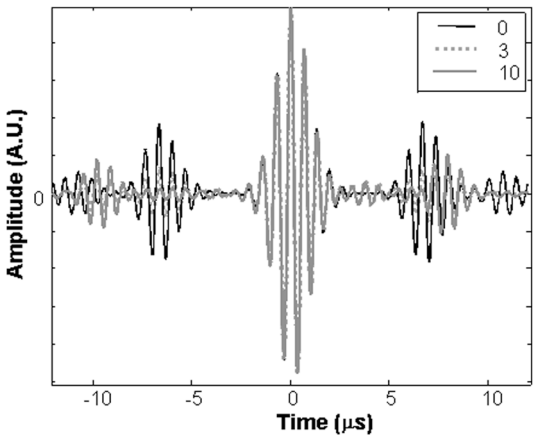
FIG. 7. Visualization of the convergence of the sequence $\mathbf{HM}_p$ for the simple iteration method. (a) Evolution of the eigenvalues of $\mathbf{HM}_p$ at its central frequency. The gray, dashed curves correspond to the value of the eigenvalues, ordered in decreasing order, for different iteration numbers (from 2 to 10). They converge to the thick black line that shows the values of the eigenvalues of the product $\mathbf{H}\hat{\mathbf{H}}^{-1}$ (inverse filter). (b) lateral repartition, in dB, of the amplitude received by array 2 when focusing on its central element using a TRM in transmission. (c) the temporal shape of the focused vector $\mathbf{HM}_pS_2$, for different values of p: plain black curve $p=0$; it corresponds to the time-reversal focusing in transmission, dashed gray curve $p=3$, plain gray curve $p=10$; it approaches the inverse filter focusing with echo cancellation and shortening of the main front duration.

B. Double iteration method

It would be of particular interest to be able to compute the inverse of $\mathbf{H}$ by using a sequence in which $\mathbf{H}$ and $\mathbf{K}_2$ do not intervene. Indeed, not having to measure $\mathbf{H}$ or $\mathbf{K}_2$ would allow us to perform noninvasive focusing: array 2 in a focal plane would no longer be needed to perform adaptive focusing. $\mathbf{H}$ can be removed from Eq. (15) by introducing a second-order sequence, replacing in Eq. (15) ${}^{t}\mathbf{H}^{*}$ by



(a)



(b)

FIG. 8. Visualization of the convergence of the sequence $\mathbf{HM}_p$ for the double iteration method. (a) evolution of the eigenvalues of $\mathbf{HM}_p$ at its central frequency. The gray, dashed curves correspond to the value of the eigenvalues, ordered in decreasing order, for different iteration orders (from 2 to 10). They converge to the thick black line that shows the values of the eigenvalues of the product $\mathbf{H}\hat{\mathbf{H}}^{-1}$ (inverse filter). (b) the temporal shape of the focused vector $\mathbf{HM}_pS_2$ for different values of p: plain black curve $p=0$; it corresponds to the time-reversal focusing in transmission; dashed gray curve $p=3$, plain gray curve $p=10$.

$$^t\mathbf{H}^* = \mathbf{M}_p - \mathbf{M}_{p-1}\mathbf{K}_2\mathbf{K}_2^*, \tag{16}$$

(15) and (16) finally lead to

$$\mathbf{M}_{p+1} = \mathbf{M}_p - \mathbf{M}_{p-1}\mathbf{K}_2\mathbf{K}_2^* + \mathbf{M}_p\mathbf{K}_2\mathbf{K}_2^*, \tag{17}$$

$$\Leftrightarrow(\mathbf{M}_{p+1} - \mathbf{M}_p) = (\mathbf{M}_p - \mathbf{M}_{p-1})\mathbf{K}_2\mathbf{K}_2^*. \tag{18}$$

It can be demonstrated (see Appendix B) that the limit $\mathbf{M}_\infty$ of this sequence is

$$\mathbf{M}_\infty = \mathbf{M}_0 + (\mathbf{M}_1 - \mathbf{M}_0)(\mathbf{H}^t\mathbf{H}^*)^{-1}. \tag{19}$$

With an adequate choice of initial values ($\mathbf{M}_0=\mathbf{0}$ and $\mathbf{M}_1 = {}^t\mathbf{H}^*$), this limit is equal to $\mathbf{H}^{-1}$.

In order to check the convergence of this sequence, $\mathbf{H}$ and $\mathbf{K}_2$ have been experimentally measured and each term $\mathbf{M}_p$ of the sequence has been successively computed thanks to (18), at all frequencies in the bandwidth of the transducers.

The evolution of the eigenvalues of ($\mathbf{HM}_p$) at central frequency is presented in Fig. 8(a). The highest eigenvalues, as with the simple iteration method, converge to 1. Another way to check the convergence is to apply the operator $\mathbf{HM}_p$ to the target signal S_2, for different values of p. For $p=0$, the vector $\mathbf{HM}_0S_2$ is equal to $\mathrm{H}^t\mathbf{H}^*S_2$, i.e., the focusing obtained by time reversal in transmission. The limit of the sequence $\mathbf{HM}_pS_2$ should be equal to $\mathbf{HH}^{-1}S_2$, i.e., the focusing by spatiotemporal inverse filter. The temporal shapes of the vectors $\mathbf{HM}_pS_2$ for $p=0,3,10$ are presented in Fig. 8(b). When iterating, the first echo preceding the main wave front and the second echo following it tend to disappear, it is not so evident for the other echoes. The second echo preceding the main wave front even seems to get slightly amplified. There is no evidence of the main wave front getting shorter. The double iteration method, being a more complicated technique to invert the propagator, seems to be less efficient and robust than the simple iteration method.

$\mathbf{K}_2$ still intervenes in Eq. (18) and $\mathbf{H}$ is still needed in the initialization of the sequence, so the double iteration method to compute $\mathbf{H}^{-1}$ still does not allow us to perform noninvasive STIF. In order to get rid of $\mathbf{K}_2$ one can work on a double iterative method to calculate the inverse of ${}^t\mathbf{H}$ instead of the inverse of $\mathbf{H}$. By analogy with Eq. (18) a matrix sequence is defined:

$$\mathbf{M}_0 = \mathbf{0},$$

$$\mathbf{M}_1 = \mathbf{H}^*,$$

$$(\mathbf{M}_{p+1} - \mathbf{M}_p) = (\mathbf{M}_p - \mathbf{M}_{p-1})\mathbf{K}_1\mathbf{K}_1^* \tag{20}$$

($\mathbf{H}$ is replaced by ${}^t\mathbf{H}^*$ and $\mathbf{K}_2$ by $\mathbf{K}_1$).

The limit $\mathbf{M}_\infty$ of this sequence is $({}^t\mathbf{H})^{-1}$ (a demonstration is the same as in Appendix B). As the transpose of the inverse is equal to the inverse of the transpose, one simply has to transpose the limit of sequence (20) to obtain the inverse of $\mathbf{H}$:

$$\mathbf{H}^{-1} = {}^t\mathbf{M}_\infty. \tag{21}$$

$\mathbf{H}$ is still needed in the initialization of the sequence (in order to reach this limit the sequence has to be initialized with $\mathbf{M}_0=\mathbf{0}$ and $\mathbf{M}_1=\mathbf{H}^*$), so the method is still invasive. Using a guess of $\mathbf{H}$ in the initialization of the sequence cannot lead to the same limit, as the limit is determined by $\mathbf{M}_1$; cf. Eq. (19):

$$\mathbf{M}_\infty = \mathbf{M}_1({}^t\mathbf{H}\mathbf{H}^*)^{-1} \text{ [(19) (modified)]}$$

[$\mathbf{M}_0$ is set to $\mathbf{0}$ and $\mathbf{H}$ is replaced by ${}^t\mathbf{H}$ in Eq. (19)].

Although it has not been possible to propose a noninvasive method for Inverse Filter focusing, two sequences have been defined to compute the inverse of a propagator by iterative methods. The matrices $\mathbf{M}_p$ of these two sequences, with p ranging from 0 to infinity, represent as many different steps between ${}^t\mathbf{H}$ and $\mathbf{H}^{-1}$. It leads to the possibility of choosing any tradeoff between the quality of the inverse filter and the robustness of the time reversal in a focusing experiment.

It should be noticed that only two of the four generalized Stokes equations have been used to derive the sequences defined by (15), (18), and (20). Information is also contained in the two remaining Stokes equations. Further work should

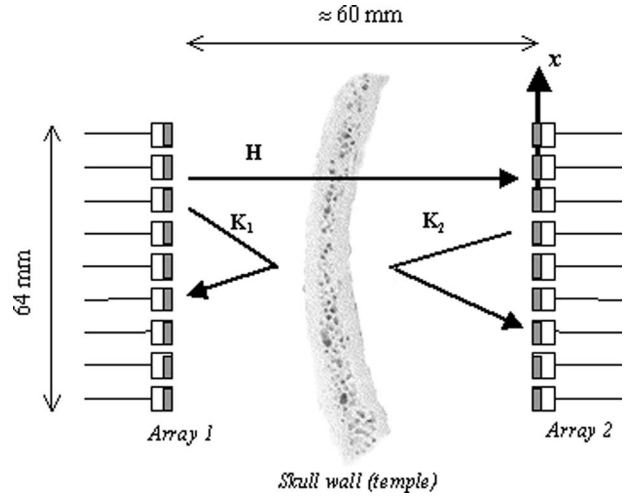


FIG. 9. A description of the skull experiment. A skull bone (temple) is inserted between the two arrays of the previous experiments instead of the solid plate. Spatial focusing in the direction x in the plane of array 2 is investigated. The two reflection propagators, from array 1 to itself $[\mathbf{K}_1(\omega)]$ and from array 2 to itself $[\mathbf{K}_2(\omega)]$, and the transmission propagator from array 1 to array 2 $[\mathbf{H}(\omega)]$, are acquired.

concentrate on all four Stokes equations to try to define a system of sequences that could hopefully lead to the noninvasive calculation of $\mathbf{H}^{-1}$.

VI. ITERATIVE INVERSION OF A PROPAGATOR THROUGH A HUMAN SKULL: APPLICATION TO SPATIAL FOCUSING

In this section the theoretical considerations made in the previous sections are applied to a more complicated object: a human skull bone. The skull bone is not planar but has a mean radius of curvature of 12 cm. It is a reflective, refractive, and absorbing heterogeneous medium. The speed of sound inside the skull ranges from 1.6 to 2.9 mm μs^{-1}.[17]

Transmission $\mathbf{H}(\omega)$ and reflection operators $\mathbf{K}_1(\omega)$ and $\mathbf{K}_2(\omega)$ are experimentally acquired through the temple of a cleaned and degased skull bone immersed in water (see a visual description of the new experimental setup in Fig. 9). $\mathbf{H}(\omega)$ is inverted thanks to the simple iteration method [Eq. (15)], the inverse of $\mathbf{H}(\omega)$ can then be used to perform spatial focusing through the skull.

Here each term $\mathbf{M}_p$ of the sequence defined by (15) using the propagators $\mathbf{H}$ and $\mathbf{K}_2$ corresponding to the skull has been successively computed at all frequencies within the bandwidth of the transducers. The evolution of the eigenvalues of $(\mathbf{HM}_p)$ at central frequency is presented in Fig. 10(a). Once again, it can be checked that the significant eigenvalues converge to 1.

Another way to check the convergence is to apply the operator $\mathbf{HM}_p$ to the new target signal S_2 defined as follows: $S_2(\omega) = {}^t[0,0,\ldots,0,1,0,\ldots,0]$: the target is now a spatiotemporal Dirac at the central position on array 2. Note that the aim is now to obtain a good *spatial* focusing, as the presence of the skull bone degrades more the spatial than the temporal focusing.[10]

The lateral amplitude distribution of the focused vectors $\mathbf{HM}_p S_2$ in the focal plane have been represented in a dB scale in Fig. 10(b) for $p=0$ (corresponding to the focusing by time reversal in transmission), $p=6$, $p=20$. The width of the main lobe when focusing with the TRM in transmission is around 3 mm: the number of resolved focal spots that can be obtained on array 2, whose lateral dimension is 64 mm, is thus 64 mm/3 mm=21; this corresponds to the number of observed significant eigenvalues [see Fig. 10(a)].

When iterating, the amplitude distribution in the focal plane gets closer to the amplitude distribution obtained with the STIF [the plain gray curve in Fig. 10(b)], in particular the sidelobe level decreases when more iterations are performed until reaching the STIF sidelobe level, up to 10 dB below the TR sidelobe level. Such results tend to demonstrate that the iterative inversion developed in previous sections can be used in a wide range of heterogeneous media.

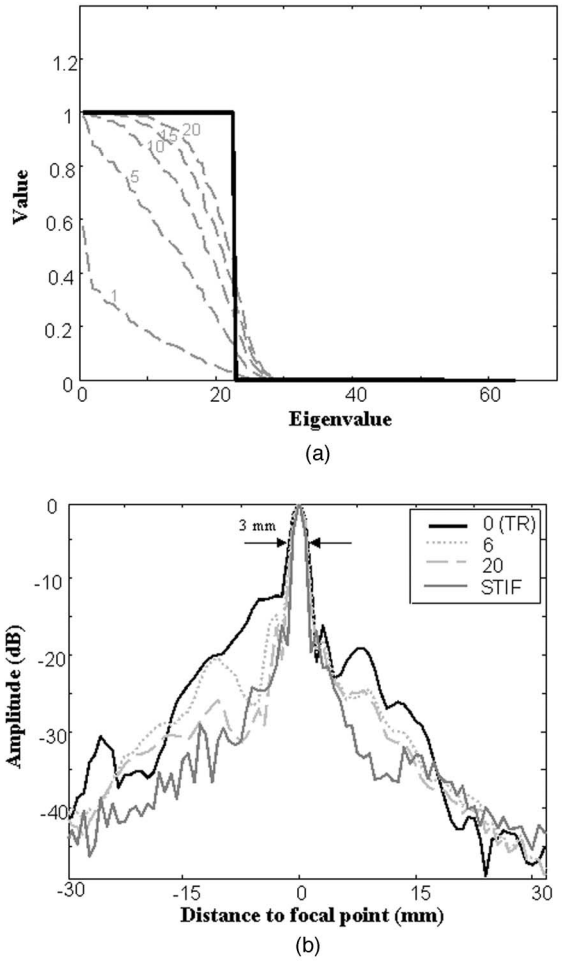


FIG. 10. Visualization of the convergence of the sequence $\mathbf{HM}_p$ for the simple iteration method, with propagators $\mathbf{H}$ and $\mathbf{K}_2$ corresponding to transmission through and reflection by a skull bone. (a) evolution of the eigenvalues of $\mathbf{HM}_p$ at its central frequency. The gray, dashed curves correspond to the value of the eigenvalues, ordered in decreasing order, for different iteration orders (from 1 to 20). They converge to the thick black line that shows the values of the eigenvalues of the product $\mathbf{H}\hat{\mathbf{H}}^{-1}$ (inverse filter). (b) Lateral amplitude repartition of the focused vector $\mathbf{HM}_p S_2$ in focal plane (S_2 is now a spatiotemporal Dirac at the central position in the focal plane) for different values of p: plain black curve, $p=0$; it corresponds to the time-reversal focusing in transmission, dotted gray curve $p=6$, and dashed gray curve $p=20$. It approaches the inverse filter focusing (plain gray curve) with a sidelobe decrease of 10 dB.

VII. CONCLUSION

It has been demonstrated that a time-reversal mirror is not sufficient to perform a good temporal compression through a plate structure because of the loss of the reflected wave front that is not recorded by the mirror. A complete time-reversal cavity made of two arrays of transducers located on each side of the plate is able to record both transmitted and reflected wave fronts, time reverse them and reemit them simultaneously, yielding to temporal compression and echo cancellation. It has been shown that echo cancellation could also be achieved by using only one array thanks to an inverse filter technique to compute the emitted set of signals. The inverse filter is able to recover the information on the reflected wave front from the only transmitted wave front. This information can indeed be recovered because transmitted and reflected waves are linked by the Stokes relations.

Thanks to a matrix formulation of the Stokes relations, the inverse filter's strategy to mimic the time-reversal cavity using only one array has been understood and highlighted. Two sequences have been proposed to compute the inverse of a propagator by an iterative method equivalent to iterative time reversal. Finally, the iterative inversion of a propagator through a human skull has been performed, showing that the theoretical considerations made in this paper are very general and can be applied to more complex media, even though they have been derived and illustrated by analyzing a simple one-dimensional experiment.

The ability of the techniques presented here to cancel the echoes induced by layered structures allow us to envision promising applications in different fields. In medical imaging, the closest application is the cancellation of reflections in bones for brain and heart imaging. In nondestructive testing, it could facilitate the detection of defects behind tubes or plates. Recently, the geological community has shown a strong interest in time reversal to reconstruct the layered structures of the earth and to detect cavities behind those structures for oil detection purposes.

In the presence of noise, the optimal focusing method lies somewhere between the robust time reversal and the powerful but noise-sensitive inverse filter. One of the iterations of the sequences presented in Sec. V can be used as a tradeoff between those two methods to reach the experimentally most efficient solution.

In summary, this work has potential implications in all applications of sound or ultrasound propagation that use arrays of emitters and receivers.

ACKNOWLEDGMENT

This work was supported by Grant No. QLG1-CT-2002-01518 (UMEDS: Ultrasonographic Monitoring and Early Diagnosis of Stroke) from the European Union.

APPENDIX A: CONVERGENCE OF THE SIMPLE ITERATION SEQUENCE

It is demonstrated here that the matrix sequence $\{\mathbf{M}_p\}$, defined by

$$\mathbf{M}_0 = \mathbf{0},$$

$$\mathbf{M}_{p+1} = {}^t\mathbf{H}^* + \mathbf{M}_p\mathbf{K}\mathbf{K}^*, \tag{A1}$$

converges to $\mathbf{H}^{-1}$ (for the sake of simplicity in this appendix the subscript 2 of $\mathbf{K}_2$ is not explicited).

Let us introduce the matrix sequence $\{\mathbf{N}_p\}$ defined by $\mathbf{N}_p = \mathbf{M}_p - \mathbf{C}$, with $\mathbf{C}$ a constant matrix that will be made explicit later, and study the convergence of this sequence.

From (13) it comes that

$$\mathbf{N}_{p+1} + \mathbf{C} = (\mathbf{N}_p + \mathbf{C})\mathbf{K}\mathbf{K}^* + {}^t\mathbf{H}^*$$

$$\Leftrightarrow \mathbf{N}_{p+1} = \mathbf{N}_p\mathbf{K}\mathbf{K}^* + [{}^t\mathbf{H}^* + \mathbf{C}(\mathbf{K}\mathbf{K}^* - \mathbf{I})]$$

$$\Leftrightarrow \mathbf{N}_{p+1} = \mathbf{N}_p\mathbf{K}\mathbf{K}^* + {}^t\mathbf{H}^* - \mathbf{C}\mathbf{H}^t\mathbf{H}^* \tag{A2}$$

$\{\mathbf{N}_p\}$ is a geometrical sequence of reason $\mathbf{K}\mathbf{K}^*$ if ${}^t\mathbf{H}^* - \mathbf{C}\mathbf{H}^t\mathbf{H}^* = \mathbf{0}$, that is, if $\mathbf{C} = {}^t\mathbf{H}^*(\mathbf{H}^t\mathbf{H}^*)^{-1} = \mathbf{H}^{-1}$. In that case

$$\mathbf{N}_p = \mathbf{N}_0(\mathbf{K}\mathbf{K}^*)^p = -\mathbf{C}(\mathbf{K}\mathbf{K}^*)^p$$

$$\Leftrightarrow \mathbf{N}_p = \mathbf{N}_0(\mathbf{I} - \mathbf{H}^t\mathbf{H}^*)^p = -\mathbf{C}(\mathbf{I} - \mathbf{H}^t\mathbf{H}^*)^p. \tag{A3}$$

As $\mathbf{M}_p = \mathbf{N}_p + \mathbf{C}$ it comes that

$$\mathbf{M}_p = \mathbf{C}[\mathbf{I} - (\mathbf{I} - \mathbf{H}^t\mathbf{H}^*)^p]$$

$$\Leftrightarrow \mathbf{M}_p = \mathbf{H}^{-1}[\mathbf{I} - (\mathbf{I} - \mathbf{H}^t\mathbf{H}^*)^p] \tag{A4}$$

($\mathbf{C}$ has been replaced by the value chosen for it, $\mathbf{H}^{-1}$).

This sequence indeed converges to $\mathbf{H}^{-1}$ as the eigenvalues of $\mathbf{I} - \mathbf{H}^t\mathbf{H}^*$ are all less than 1: $\mathbf{H}^t\mathbf{H}^*$ and $\mathbf{K}\mathbf{K}^*$ are both hermitic then diagonalizable with all positive eigenvalues, and the Stokes relation (S2') ($\mathbf{H}^t\mathbf{H}^* + \mathbf{K}\mathbf{K}^* = \mathbf{I}$) implies that the eigenvalues of both $\mathbf{H}^t\mathbf{H}^*$ and $\mathbf{K}\mathbf{K}^* = \mathbf{I} - \mathbf{H}^t\mathbf{H}^*$ are all between 0 and 1.

APPENDIX B: CONVERGENCE OF THE DOUBLE ITERATION SEQUENCE

It is demonstrated here that the matrix sequence $\{\mathbf{M}_p\}$, defined by

$$\mathbf{M_0} = \mathbf{0},$$

$$\mathbf{M}_1 = {}^t\mathbf{H}^*,$$

$$(\mathbf{M}_{p+1} - \mathbf{M}_p) = (\mathbf{M}_p - \mathbf{M}_{p-1})\mathbf{K}\mathbf{K}^* \tag{B1}$$

converges to $\mathbf{H}^{-1}$ (for the sake of simplicity in this appendix the subscript 2 of $\mathbf{K}_2$ is not explicited).

$\mathbf{K}\mathbf{K}^*$ is hermitic then diagonalizable with positive eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. Sequences $\{\mathbf{W}_p\}$ and $\{\mathbf{U}_p\}$ are defined by

$$\mathbf{W}_p = \mathbf{M}_{p+1} - \mathbf{M}_p, \tag{B2}$$

$$\mathbf{U}_p = \sum_{k=0}^{p} \mathbf{W}_k. \tag{B3}$$

This implies that

$$\mathbf{U}_p = \mathbf{M}_{p+1} - \mathbf{M}_0. \tag{B4}$$

From (B1) and (B2) it comes that

$$\mathbf{W}_{p+1} = \mathbf{W}_p \mathbf{K}\mathbf{K}^* \tag{B5}$$

$$\Rightarrow \mathbf{W}_p = \mathbf{W}_0 (\mathbf{K}\mathbf{K}^*)^p = \mathbf{W}_0 \begin{pmatrix} \lambda_1^p & & & \\ & \lambda_2^p & 0 & \\ & & \ddots & \\ & 0 & & \lambda_n^p \end{pmatrix}, \tag{B6}$$

in the base of the eigenvectors.

Then,

$$\mathbf{U}_p = \sum_{k=0}^{p} \mathbf{W}_k = \mathbf{W}_0 \begin{pmatrix} 1+\lambda_1+\cdots+\lambda_1^p & & & \\ & 1+\lambda_2+\cdots+\lambda_2^p & 0 & \\ & & \ddots & \\ & 0 & & 1+\lambda_n+\cdots+\lambda_n^p \end{pmatrix}, \tag{B7}$$

$$\Leftrightarrow \mathbf{U}_p = \mathbf{W}_0 \begin{pmatrix} (1-\lambda_1^p)/(1-\lambda_1) & & & \\ & (1-\lambda_2^p)/(1-\lambda_2) & 0 & \\ & & \ddots & \\ & 0 & & (1-\lambda_n^p)/(1-\lambda_n) \end{pmatrix}. \tag{B8}$$

The limit of $\{\mathbf{U}_p\}$ is then, as all the eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ of $\mathbf{K}\mathbf{K}^*$ are between 0 and 1 (see Appendix A for a demonstration),

$$\mathbf{U} = \mathbf{W}_0 \begin{pmatrix} 1(/1-\lambda_1) & & & \\ & 1/(1-\lambda_2) & 0 & \\ & & \ddots & \\ & 0 & & 1/(1-\lambda_n) \end{pmatrix}. \tag{B9}$$

From the first Stokes relation (S1') one knows that $\{1-\lambda_1, 1-\lambda_2, \ldots, 1-\lambda_n\}$ are the eigenvalues of $\mathbf{H}^t\mathbf{H}^*$, and then

$$\mathbf{U} = \mathbf{W}_0 (\mathbf{H}^t\mathbf{H}^*)^{-1} \tag{B10}$$

From (B3), (B4), and (B10), it comes then that the sequence $\{\mathbf{M}_p\}$ converges, and its limit is

$$\mathbf{M} = \mathbf{M}_0 + (\mathbf{M}_1 - \mathbf{M}_0)(\mathbf{H}^t\mathbf{H}^*)^{-1}. \tag{B11}$$

[1] M. Fink, "Time reversal of ultrasonic fields-Part I: Basic principles," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **39**, 555–566 (1992).

[2] R. G. Sachs, *The Physics of Time Reversal* (The University of Chicago Press, Chicago, 1987).

[3] L. Borcea, G. Papanicolaou, C. Tsoga, and J. Berryman, "Imaging and time reversal in random media" Inverse Probl. Eng. **18**, 1247–1279 (2002).

[4] A. Derode, P. Roux, and M. Fink, "Robust acoustic time reversal with high-order multiple scattering," Phys. Rev. Lett. **75**, 4206–4209 (1995).

[5] D. Cassereau and M. Fink, "Focusing with plane time-reversal mirrors: an efficient alternative to closed cavities," J. Acoust. Soc. Am. **94**, 2373–2386 (1993).

[6] E. Ebbini and C. Cain, "Multiple-focus ultrasound phased-array pattern synthesis: optimal driving-signal distributions for hyperthermia," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **36**, 540–548 (1989).

[7] H. Wang, E. Ebbini, M. O'Donnell, and C. Cain, "Phase aberration correction and motion compensation for ultrasonic hyperthermia phased arrays: Experimental results," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **41**, 34–43 (1994).

[8] E. Ebbini and J. Mok-Kun, "Optimal synthesis of coded wavefronts for ultrasonic pulse-echo imaging," *Proc. SPIE Vol. 4325*, Medical Imaging 2001, pp. 94–101.

[9] M. Tanter, J.-F. Aubry, J. Gerber, J.-L. Thomas, and M. Fink, "Optimal focusing by spatio temporal inverse filter. Part I: Basic Principles," J. Acoust. Soc. Am. **110**, 37–47 (2001).

[10] J.-F. Aubry, M. Tanter, J. Gerber, J.-L. Thomas, and M. Fink, "Optimal focusing by spatio temporal inverse filter. II. Application to focusing through absorbing and reverberating media," J. Acoust. Soc. Am. **110**, 48–58 (2001).

[11] M. Tanter, J. L. Thomas, and M. Fink, "Time reversal and the inverse filter," J. Acoust. Soc. Am. **108**, 223–234 (2000).

[12] G. Montaldo, M. Tanter, and M. Fink, "Real time inverse filter through iterative time reversal," J. Acoust. Soc. Am. **115**, 768 (2004).

[13] D. Cassereau and M. Fink, "Time reversal focusing through a plane interface separating two fluids," J. Acoust. Soc. Am. **96**, 3145–3154 (1994).

[14] C. Tsoga and G. Papanicolaou, "Time reversal through a solid–liquid interface and super-resolution," Inverse Probl. Eng. **18**, 1639–1657 (2002).

[15] M. Nieto-Vesperinas and E. Wolf, J. Opt. Soc. Am. A **2**, 1429 (1985).

[16] M. Nieto-Vesperinas, *Scattering and Diffraction in Physical Optics* (Wiley, New York, 1991).

[17] F. Fry and J. Barger, "Acoustical properties of the human skull," J. Acoust. Soc. Am. **63**, 1576–1590 (1978).

Dispersive surface waves along partially saturated porous media

Gabriel Chao[a)]
Department of Geotechnology, Delft University of Technology, and Department of Applied Physics, Eindhoven University of Technology, P. O. Box 513, 5600 MB Eindhoven, The Netherlands

D. M. J. Smeulders
Department of Geotechnology, Delft University of Technology, P. O. Box 5028, 2600 6A Delft, The Netherlands

M. E. H. van Dongen
Department of Applied Physics, Eindhoven University of Technology, P. O. Box 513, 5600 MB Eindhoven, The Netherlands

(Received 7 July 2005; revised 2 December 2005; accepted 13 December 2005)

Numerical results for the velocity and attenuation of surface wave modes in fully permeable liquid/partially saturated porous solid plane interfaces are reported in a broadband of frequencies (100 Hz–1 MHz). A modified Biot theory of poromechanics is implemented which takes into account the interaction between the gas bubbles and both the liquid and the solid phases of the porous material through acoustic radiation and viscous and thermal dissipation. This model was previously verified by shock wave experiments. In the present paper this formulation is extended to account for grain compressibility. The dependence of the frequency-dependent velocities and attenuation coefficients of the surface modes on the gas saturation is studied. The results show a significant dependence of the velocities and attenuation of the pseudo-Stoneley wave and the pseudo-Rayleigh wave on the liquid saturation in the pores. Maximum values in the attenuation coefficient of the pseudo-Stoneley wave are obtained in the 10–20 kHz range of frequencies. The attenuation value and the characteristic frequency of this maximum depend on the liquid saturation. In the high-frequency limit, a transition is found between the pseudo-Stoneley wave and a true Stoneley mode. This transition occurs at a typical saturation below which the slow compressional wave propagates faster than the pseudo-Stoneley wave. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2164997]

PACS number(s): 43.20.Jr, 43.20.Gp [RR] Pages: 1347–1355

I. INTRODUCTION

The presence of gas bubbles can dramatically influence the acoustic properties of a liquid. The bulk modulus of the liquid becomes frequency-dependent and attenuation effects arise due to oscillations of the bubbles (radiation) and heat transfer to the surrounding liquid.[1] It is particularly interesting to consider the problem of a gas-liquid mixture filling the pore space of a porous medium. In this case, even more dissipative mechanisms have to be taken into account, namely the interaction between the gas and both the liquid and the solid elastic matrix. In the case that only liquid saturates the pore space, the interaction between the liquid and the solid matrix can be understood in terms of the Biot theory.[2,3] This theory was previously extended in order to include the effects of gas saturation on the bulk elastic waves in partially saturated porous media by among others White,[4] Dutta and Ode,[5,6] Berryman *et al.*,[7] Smeulders and Van Dongen,[8] Johnson,[9] and Carcione *et al.*[10]

A great deal of attention has been given to the influence of the gas saturation on the velocities and attenuation of seismic waves since the pioneering work of White.[4] The White model describes the air fraction as spherical gas pockets distributed in a cubic array in the liquid-saturated porous medium. This model will be referred to as the "gas pocket model." Dutta and Ode[5,6] provided a more complete solution based on Biot's theory for the bulk modulus of a single bubble surrounded by a fluid-saturated porous spherical shell. Berryman *et al.*[7] formulated a model based on variational principles for the bulk acoustic properties of a porous medium saturated with a mixture of two fluids. Experiments were carried out by Smeulders and Van Dongen[8] on compressional wave propagation in porous columns saturated by an air-water mixture. Their theoretical model is based on the study of the response of a gas bubble in a fully saturated porous medium to an external oscillating pressure field. Damping mechanisms due to radiation into the two compressional waves, viscous dilatation at the bubble surface, and heat exchange with the solid matrix are considered. Degrande *et al.*[11] used this model to study the effects of saturation on the wave propagation phenomena in a porous layer adjacent to a water table. An interpretation of laboratory velocity measurements in a variety of partially gas-saturated rocks is given by Gist.[12] Cadoret *et al.*[13,14] reported experimental results using a resonant-bar technique to determine

a)Electronic mail: g.e.chao@tue.nl

the velocity and attenuation of acoustic waves in partially saturated limestones at a sonic frequency of 1 kHz. Similar experiments were previously performed by Lucet.[15]

Despite all the efforts and attention to study saturation effects on seismic and acoustic waves, there is, to our best knowledge, no study concerning the influence of the liquid saturation on surface waves. The purpose of this work is to investigate the effects of the gas fraction on the propagation of surface waves along a plane interface between a liquid and a partially saturated porous medium. The bulk acoustic properties of the partially saturated porous medium are described according to the model of Smeulders and Van Dongen.[8] The high-frequency properties of the surface waves for the fully saturated case were studied in detail by Feng and Johnson.[16,17] There are three surface modes that can propagate depending on the relation between the mechanical properties of the porous material and the liquid, and the characteristics of the interface regarding the possibility for the liquid to flow between the two half-spaces (surface permeability). The three modes are the Stoneley wave, the pseudo-Stoneley wave and the pseudo-Rayleigh wave. The Stoneley wave is a true surface wave which propagates almost undamped along the interface with an exponential decay in the normal direction away from the interface. The pseudomodes are significantly damped in the direction of propagation and radiate energy into the slow compressional wave only (pseudo-Stoneley wave) or both into the slow compressional wave and the acoustic wave in the liquid half-space (pseudo-Rayleigh wave). Recently, Gubaidullin *et al.*[18] considered the effects of viscous losses in the dispersive properties of the surface waves. In this paper we consider the influence of gas bubbles in the porous solid on the properties of the surface waves. First we investigate the high-frequency limit, where the viscous interaction can be neglected. Then the frequency-dependent dispersion of the pseudo modes is analyzed.

The paper is organized as follows. In Sec. II we review the theoretical model for acoustic wave propagation for the case that a liquid-gas mixture saturates the porous material. In Sec. III the results for the velocity and attenuation of the surface modes propagating along a liquid-poroelastic plane interface are presented and discussed. First the high-frequency limit is examined and the different waves are discussed, followed by the analysis of the frequency-dependent results. The study is summarized and the conclusions are given in Sec. IV.

II. ACOUSTIC PROPERTIES OF A PARTIALLY SATURATED POROUS MEDIUM

Acoustic wave propagation through a fully saturated porous media can be described in terms of the Biot equations. In the frequency domain these equations are expressed as

$$-\omega^2(\tilde{\rho}_{11}\tilde{\boldsymbol{u}}+\tilde{\rho}_{12}\tilde{\boldsymbol{U}})=(P-N)\,\boldsymbol{\nabla}\,\boldsymbol{\nabla}\cdot\tilde{\boldsymbol{u}}+N\boldsymbol{\nabla}^2\tilde{\boldsymbol{u}}+Q\,\boldsymbol{\nabla}\,\boldsymbol{\nabla}\cdot\tilde{\boldsymbol{U}} \tag{1}$$

and

$$-\omega^2(\tilde{\rho}_{12}\tilde{\boldsymbol{u}}+\tilde{\rho}_{22}\tilde{\boldsymbol{U}})=R\,\boldsymbol{\nabla}\,\boldsymbol{\nabla}\cdot\boldsymbol{U}+Q\,\boldsymbol{\nabla}\,\boldsymbol{\nabla}\cdot\tilde{\boldsymbol{u}}, \tag{2}$$

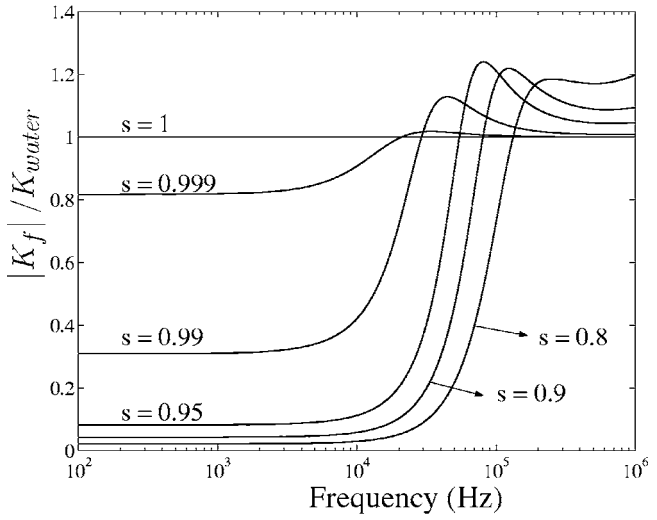
where $\tilde{\boldsymbol{u}}$ is the solid displacement and $\tilde{\boldsymbol{U}}$ is the fluid displacement. N is the shear modulus of the composite material and P, Q, and R are the so-called generalized elastic coefficients. They are related to the porosity ϕ, the solid frame bulk modulus K_b, the solid grain bulk modulus K_s, the pore-fluid modulus, K_f and N through the so-called Gedanken experiments. The parameters $\tilde{\rho}_{11}$, $\tilde{\rho}_{12}$, and $\tilde{\rho}_{22}$ are the complex-valued frequency-dependent densities. They are functions of the density of the fluid ρ_f, the density of the solid ρ_s, the porosity ϕ, and the frequency-dependent tortuosity $\tilde{\alpha}(\omega)$. It is not the purpose of this section to review Biot's theory and for further details the reader is referred to classical books on the subject (see, e.g., Allard[19] and Bourbie *et al.*[20]).

In our case, the pore space is saturated by a mixture of water and air. Therefore, new interaction mechanisms between the gas and the liquid and the gas and the solid matrix have to be taken into account. The oscillations of the air bubbles will induce radiation of the two compressional waves at the bubble surfaces. The liquid dilatation at the bubble surface causes viscous attenuation. Finally, heat transport from the bubble to the surrounding media is also considered. In this work, the dissipative phenomena mentioned above are described in terms of a complex-valued frequency-dependent bulk modulus of the mixture of water and air.[8] In this section we will review the main results of this theory and analyze its implications for the bulk modes. The compressibility of the solid grains is also considered, which was neglected in the cited paper.[8] The model is based on the calculation of the volume variation of a single bubble as a response to an external oscillating pressure field (Appendix A). The dynamics of the bubble is determined by the solution of the Biot equations at the spherical interface between the gas-saturated and the liquid-saturated porous media. Mathematically, it is possible to solve the Biot equations in spherical coordinates in the two domains, inside and outside the bubble. The solutions are then matched using appropriate boundary conditions and the bubble volume change due to the harmonic pressure can be calculated. In this way the bulk modulus of the bubble can be computed and, neglecting the interaction between the bubbles it will be considered as the bulk modulus of the gas phase in the mixture, $K_g(\omega)$. The frequency-dependent bulk modulus of the mixture, $K_f(\omega)$ is obtained through a modified Wood's formula[21]

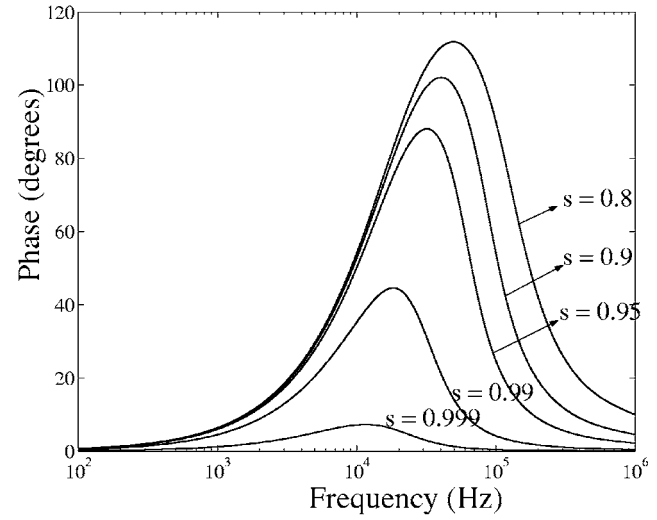
$$\frac{1}{K_f(\omega)}=\frac{s}{K_l}+\frac{1-s}{K_g(\omega)}, \tag{3}$$

where K_l is the bulk modulus of the liquid phase and s is the liquid saturation. The expression for $K_f(\omega)$ given in Eq. (3) differs from the original Wood's formula in which both the bulk modulus for the gas and liquid phases are constant.

Strictly speaking, the original Wood's formula is only valid for highly homogeneous mixtures and at frequencies sufficiently low so that the wavelengths are considerably larger than the size of the heterogeneities. In this case it is possible to assume that the external oscillating pressure field is spatially homogeneous at a local scale. In our case this scale is determined by the size of the gas bubbles and the



(a)



(b)

FIG. 1. Frequency-dependent bulk modulus for a mixture of water and air saturating a Berea sandstone porous rock. The radius of the air bubbles is 1 mm and the gas pressure is 0.01 GPa (100 bars). Different liquid saturation s are considered.

distance between them. This assumption is valid for the range of bubble sizes and frequencies considered in this work. At higher frequencies or heterogeneous mixtures, scattering effects cannot be neglected and it is no longer possible to define a homogeneous external driving pressure at a local scale. In this work a bubble radius of 1 mm is considered, in accordance with experimental values reported for air-water mixtures saturating the pores of artificial sandstones.[8] In this scenario, a threshold frequency of 1.5 MHz can be defined, below which the assumption of this model is valid. At this threshold frequency, the wavelength of the fast compressional wave equals the diameter of the gas bubble.

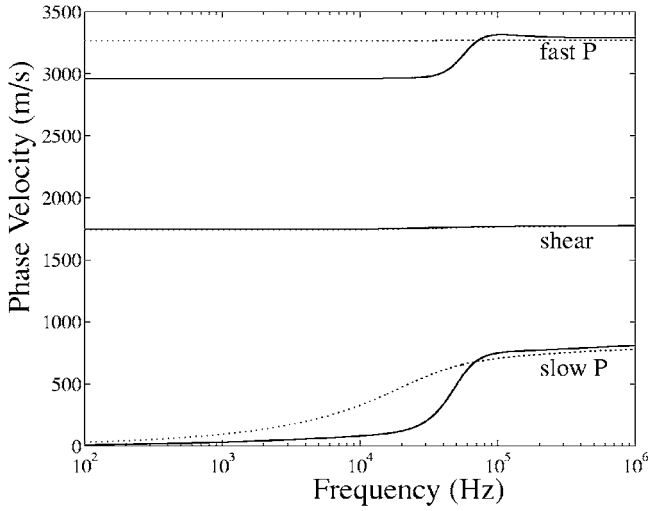
Figure 1 shows the absolute and phase values of the bulk modulus of the mixture as a function of the frequency for different liquid saturations. A Berea sandstone saturated by a water-air mixture is considered. The properties of the porous material and the saturating fluids are given in Table I. On one hand, at low frequencies, the bulk modulus of the gas phase equals 0.01 GPa and therefore a decrease in liquid saturation causes a decrease in the bulk modulus of the mixture since $K_l > K_g$. On the other hand, at high frequencies, the gas phase becomes highly incompressible ($|K_g| \to \infty$) and $K_f = K_l/s$. In this limit, K_f increases with concentration of air in the water. The transition between the two limits shows a minimum in the compressibility of the mixture, which corresponds to the antiresonance frequency of the bubble. At this frequency the bubbles oscillate out-of-phase with the external pressure field, which results in a highly incompressible medium. The relevant parameters in this model are the pressure of the gas, its saturation in the pore space, and the radius of the gas bubbles. The outcome of the velocities and attenuation of the compressional waves that propagate in this partially saturated porous media are shown in Fig. 2 for a liquid saturation s of 0.95 and a bubble radius of 1 mm. The results for the wave velocities can be explained by the arguments about the changes in the compressibility of the mixture discussed above. The presence of air decreases the bulk modulus of the mixture at low frequencies which results in compressional waves propagating slower in the partially saturated case. This behavior is reversed at high frequencies where the compressional waves propagate faster when the air saturation is increased. More interesting are the modifications induced by the air phase in the attenuation coefficients. The decrease in the liquid content of the mixture result in a significant increase of the attenuation for the fast compressional wave, which is observed throughout the complete range of frequencies studied. The slow compressional wave presents a maximum in the attenuation for the partially saturated case. This maximum is not observed for the fully saturated case. The model presented here assumes that the shear wave is influenced by the presence of the gas phase only due to changes in density. The frequency-dependent mechanisms incorporated in this model have been experimentally corroborated by shock-induced transmission/reflection wave experiments carried out in a shock tube.[8]

TABLE I. Physical properties of the Berea sandstone and the saturating fluids: water and air.

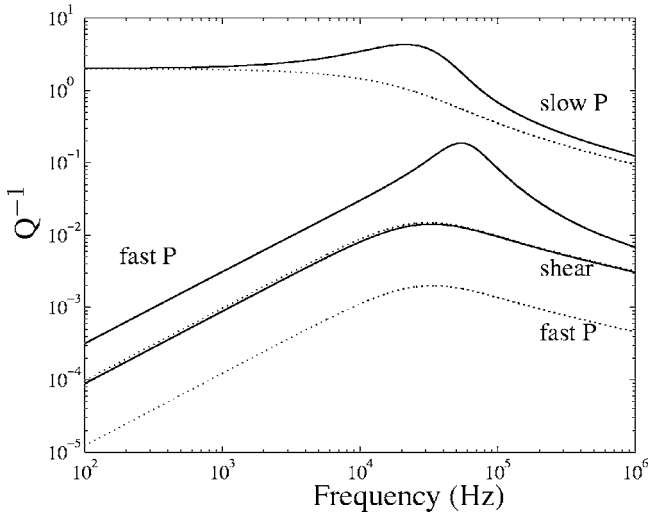
Solid density ρ_s (kg/m^3)	2644
Porosity ϕ	0.20
Permeability k_0 (mD)	360
Tortuosity α_∞	2.4
Frame bulk modulus K_b (GPa)	10.37
Shear modulus N (GPa)	7.02
Grain bulk modulus K_s (GPa)	36.5
Liquid bulk modulus K_l (GPa)	2.25
Gas pressure (bulk modulus) p_g (GPa)	0.01
Liquid density ρ_l (kg/m^3)	1000
Gas density ρ_g (kg/m^3)	100
Liquid viscosity η_l (mPa s)	1
Gas viscosity η_g (mPa s)	1.5×10^{-2}
Gas thermal diffusivity a_g	1.8×10^{-7}

III. SATURATION EFFECTS ON THE VELOCITIES AND ATTENUATION OF THE SURFACE WAVES

In this section the numerical results for the phase velocities and attenuation coefficients of the surface wave modes



(a)



(b)

FIG. 2. Phase velocities (a) and attenuation coefficients (b) of the body waves in a water/air-saturated Berea sandstone. The effects of air saturation are shown for the compressional waves in solid lines, the gas pressure is 0.01 GPa, the bubble radius 1 mm, and $s=0.95$. The Biot predictions for the fully saturated case are shown in dotted lines.

that propagate along a liquid/partially saturated poroelastic plane interface are discussed. The configuration is depicted in Fig. 3. The mathematical procedure involves the numerical solution of the boundary value problem which follows

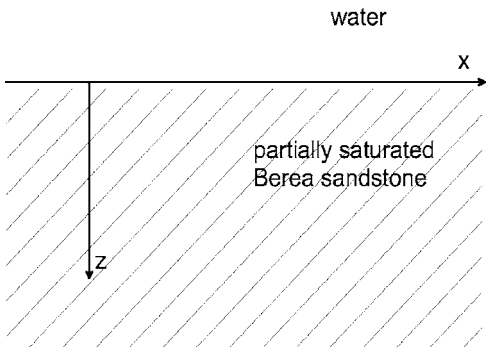
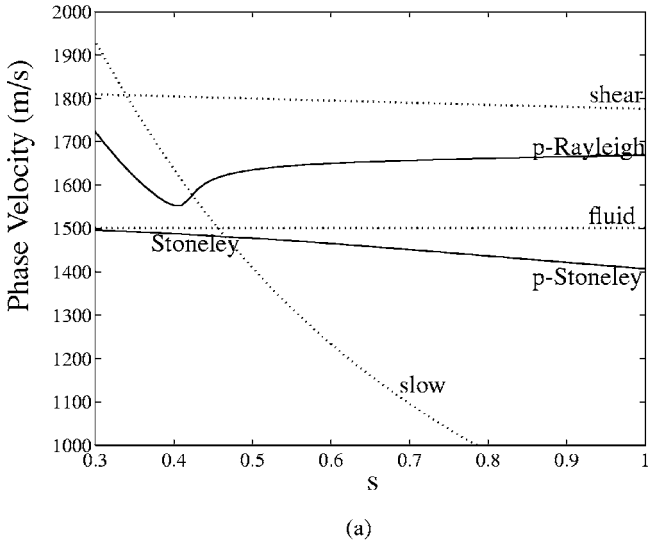


FIG. 3. Liquid/partially saturated porous medium plane interface.

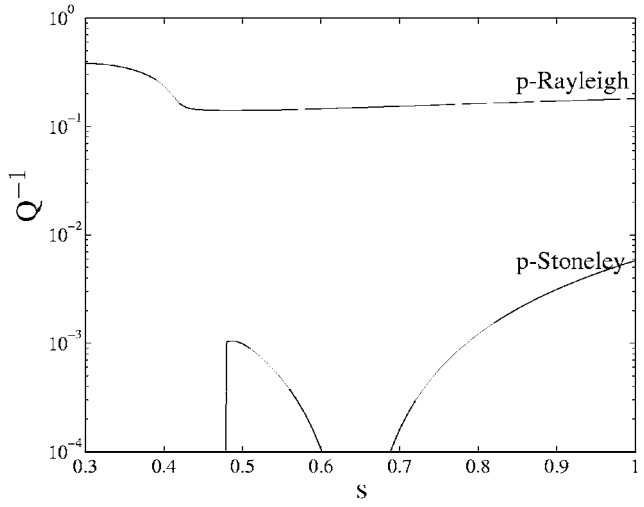
from the application of the boundary conditions at the interface (see Appendix B). The oscillating gas bubble model[8] outlined in the previous section is employed to describe the bulk modulus of the fluid phase, which in this case is composed of a mixture of water and air. The properties of the solid matrix correspond to the Berea sandstone characterized in Table I. We adopt the surface wave terminology given by Feng and Johnson.[16] In order to avoid confusion, it is worthwhile to mention that the pseudo-Stoneley wave propagating along a liquid/poroelastic interface is the generalization of the classical Stoneley wave in a liquid/elastic interface. In the poroelastic case it becomes a pseudo wave due to radiation into the slow P wave. It is important to note that in this work we will assume that the interface is fully permeable so that continuity of pressure holds across the interface. The effect of sealed or partially sealed pores at the interface has been modelled in the past using the empirical concept of surface flow impedance. We restrict ourselves to the open pore boundary case.

We first examine the high-frequency limit, for which the velocities of the bulk modes become real valued and the slow wave is propagative. It also holds that the bulk modulus of the mixture saturating the pore space becomes real-valued (K_l/s). Therefore, the dissipative mechanisms induced by the oscillation of the gas bubbles are not present in this limit as can be clearly observed in Fig. 1. The dependence of the surface wave velocities and attenuation on the water saturation is shown in Fig. 4. For reference, the bulk wave velocities are also displayed.

For the fully water-saturated case $s=1$, two surface modes are found: the pseudo-Stoneley wave and the pseudo-Rayleigh wave. The pseudo-Stoneley wave has a velocity which is faster than the velocity of the slow wave and slower than the speed of the rest of the bulk modes. This implies that it radiates energy into the slow wave and therefore it is called a pseudo or leaky mode. The pseudo-Rayleigh wave leaks energy into the fluid half-space and into the slow wave, its velocity is faster than that of the slow wave and the fluid wave but slower than that of the shear and the fast wave (the fast wave is not plotted). The velocity of the slow wave decreases with increasing water saturation as can be observed in Fig. 4, while the shear mode speed is slightly affected due to density effects only. The behavior of the slow wave as a function of saturation and its relation with the other bulk modes plays an important role in the properties of the surface waves. For water saturations higher than 0.47, the velocity of the pseudo-Stoneley wave is higher than that of the slow wave. In this range of saturations both the pseudo-Stoneley wave and the pseudo-Rayleigh wave exist. The velocity of the pseudo-Stoneley wave decreases with increasing water saturation. For s values below 0.47 the pseudo-Stoneley wave becomes a true Stoneley wave due to the fact that the slow wave becomes faster than it. This transition is neatly illustrated in the attenuation coefficient [Fig. 4(b)], which shows the attenuation in terms of the inverse quality factor Q^{-1}. When the pseudo-Stoneley wave becomes the true Stoneley wave, the damping necessarily disappears because radiation ceases to exist for saturation values below 0.47. The attenuation of the pseudo-Stoneley wave has a

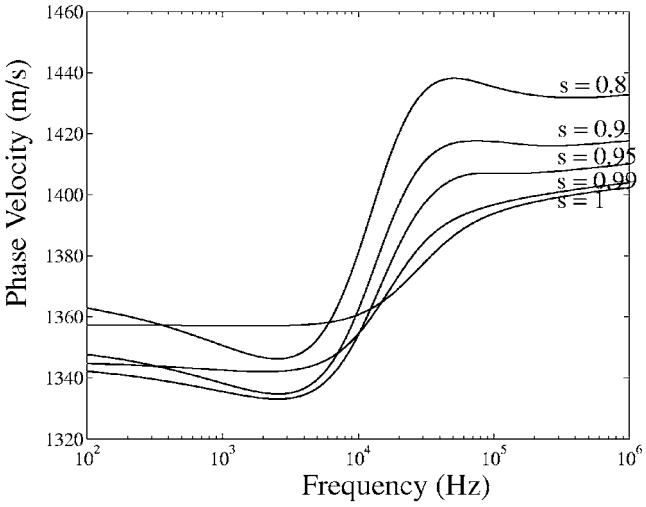


(a)

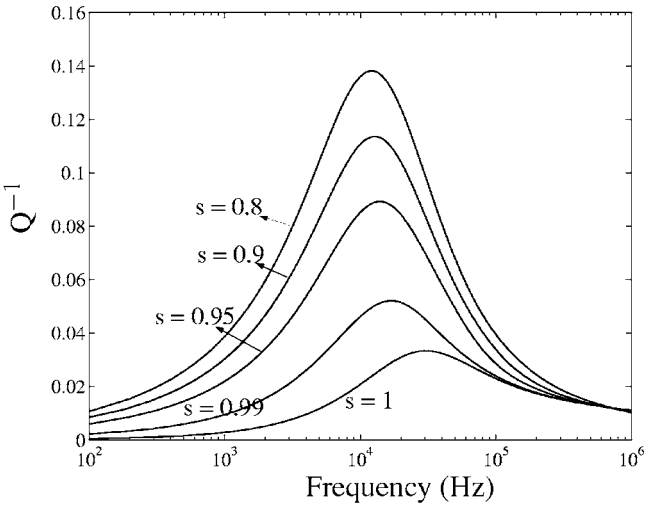


(b)

FIG. 4. Saturation effects on the phase velocity (a) and attenuation (b) of the surface waves that propagate in a flat interface between water and a porous Berea sandstone saturated by a mixture of air and water. The high-frequency limit is considered. The bulk wave velocities are plotted in dashed lines.



(a)



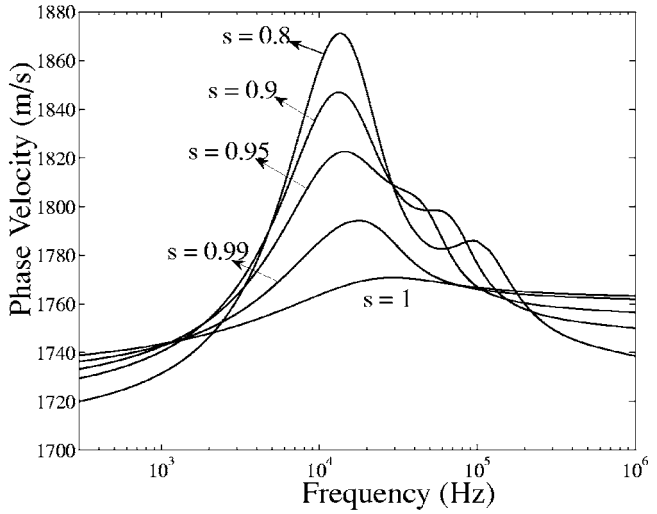
(b)

FIG. 5. Frequency-dependent phase velocities (a) and attenuation coefficients (b) of the pseudo-Stoneley wave along a water/partially saturated poroelastic interface. The porous material is a Berea sandstone saturated with a water-air mixture. Different liquid saturations s are considered.

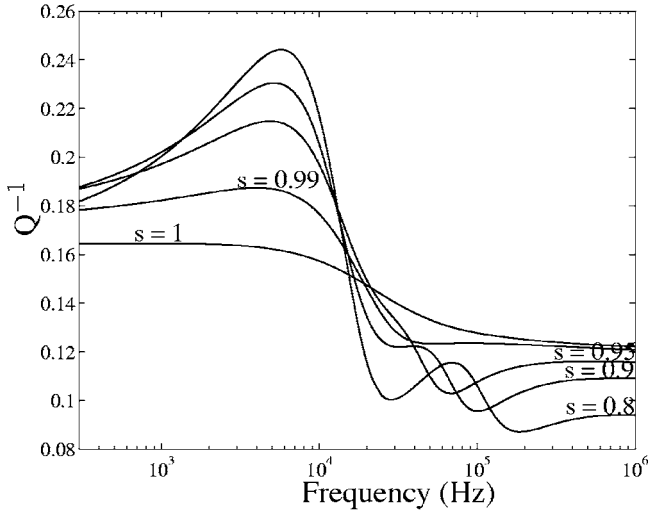
sharp minimum at 0.64 and over the entire range of saturations it is significantly less damped than the pseudo-Rayleigh mode. The pseudo-Rayleigh wave ceases to radiate into the slow wave for water saturations below 0.43, because its velocity becomes lower than that of the slow wave. A sharp increase in Q^{-1} is observed for water saturations below 0.43 where the pseudo-Rayleigh wave ceases to radiate into the slow wave. The only dissipative mechanism here is radiation into the fluid wave. It is worthwhile to note that the study of the high-frequency limit provides a first insight on the compressibility effects on the surface modes due to the presence of the gas fraction.

We now extend the study to more realistic frequency-dependent surface waves and we consider the dissipative mechanisms which were neglected previously. In this case we calculate the dispersive results for the leaky modes for different liquid saturations. Figure 5 shows the results for the pseudo-Stoneley wave. The phase velocity decreases with saturation at high frequencies. This is consistent with the high-frequency results. At low frequencies, the phase velocities are considerably less than at high frequencies for all saturations, although a clear trend with saturation is not found. The results for the attenuation coefficient show a monotonous increase with the air fraction occupying the pore space. This influence of the gas fraction is most significant in the 1–100 kHz range. The characteristic frequency for the maximum of Q^{-1} depends on saturation. It is found that when s increases this characteristic frequency moves towards higher values.

The influence of the saturation on the properties of the pseudo-Rayleigh is depicted in Fig. 6. At low and high frequencies the pseudo-Rayleigh wave propagates slower when the liquid fraction in the pore space is decreased. Interesting features occur at intermediate frequencies (1–150 kHz). In this range of frequencies the speed of the pseudo-Rayleigh mode decreases with increasing saturation. Furthermore, a



(a)



(b)

FIG. 6. Frequency-dependent phase velocities (a) and attenuation coefficients (b) of the pseudo-Rayleigh wave in a water/partially saturated poroelastic interface. The porous material is a Berea sandstone saturated with a water-air mixture. Different liquid saturations s are considered.

peak in the phase velocity is predicted. For the values of liquid saturation studied in this work, this maximum lies in frequencies between 10 and 20 kHz. The position of this maximum on the frequency axis slightly depends on saturation; lower characteristic frequencies are obtained for lower values of s. It is interesting to note the presence of additional local maxima for the $s=0.95$, $s=0.9$, and $s=0.8$ cases, which become more pronounced for lower saturation values. The characteristic frequency of this secondary maximum increases with decreasing saturation.

The higher attenuation values are obtained in the low-frequency range and a maximum is observed. This maximum is associated with the presence of air bubbles and becomes sharper when the saturation decreases. For the lower liquid saturation cases considered here, $s=0.9$ and $s=0.8$, a second local maximum is observed at higher frequencies. The attenuation coefficient Q^{-1} diminishes at high frequencies and in this limit a clear dependence on liquid saturation is found where the attenuation increases with saturation.

IV. CONCLUSIONS AND DISCUSSION

In this work we have studied the saturation effects on the properties of the surface waves that propagate along a plane interface between a liquid and a partially saturated porous solid. The numerical results for the pseudo-Stoneley wave and the pseudo-Rayleigh wave show interesting features when the pore space of the poroelastic medium is filled with a mixture of water and air. In the high-frequency limit where only compressibility effects are present, the full range of liquid saturations was studied. A transition between the leaky pseudo-Stoneley wave and the true Stoneley wave is found at a characteristic saturation for which the slow P wave propagates faster than the pseudo-Stoneley wave. This transition is neatly illustrated in the behavior of the attenuation coefficient Q^{-1} which drastically decreases for water saturations lower than $s=0.46$. This indicates that the pseudo-Stoneley wave becomes a true unattenuated surface wave, the Stoneley wave.

When the frequency-dependent dissipative mechanisms are included, interesting features arise in the velocity and attenuation of the surface waves. The pseudo-Stoneley wave shows a well-defined maximum in the attenuation. This maximum is located in the range of frequencies which is relevant to borehole geophysical applications (5–30 kHz). The characteristic frequency of this maximum depends on the liquid saturation. In acoustic borehole logging techniques the pseudo-Stoneley plays an important role in reservoir characterization. In this context, our numerical results indicate that the attenuation of the pseudo-Stoneley can provide valuable information on the liquid saturation in the pores. Similar conclusions can be drawn for the phase velocity of the pseudo-Rayleigh wave though it should be noted that this wave is difficult to detect in field or laboratory measurements. An independent determination of the bubble size remains the main obstacle for a direct application of this model to practical situations.

ACKNOWLEDGMENT

This study was supported by the ISES (Integrated Solid Earth Sciences) program.

APPENDIX A: FREQUENCY-DEPENDENT BULK MODULUS $K_g(\omega)$

The purpose of this appendix is to highlight the main conceptual steps involved in the derivation of the complex-valued bulk modulus of the gas phase, $K_g(\omega)$. In the remaining of the appendixes and in order to simplify the notation, the tilde above the functions and quantities in the frequency domain is omitted. The tilde above the density terms and the tortuosity is used to denote the frequency-dependent nature of these functions (see, e.g., Allard[19]).

Let us consider a spherical air bubble immersed in a fully water saturated porous medium in the presence of an external oscillating pressure field. First, we will focus on the

external domain (fully water-saturated porous medium outside the bubble). We introduce the displacement potentials Φ_{c1} and Φ_{c2} associated with the fast wave and the slow compressional wave as follows:

$$\boldsymbol{u} = \boldsymbol{\nabla}\Phi_{c1} + \boldsymbol{\nabla}\Phi_{c2}, \tag{A1}$$

and

$$\boldsymbol{U} = G_{c1}\,\boldsymbol{\nabla}\,\Phi_{c1} + G_{c2}\,\boldsymbol{\nabla}\,\Phi_{c2}, \tag{A2}$$

where

$$G_{c1} = \frac{P - v_{c1}^2\tilde{\rho}_{11}}{v_{c1}^2\tilde{\rho}_{12} - Q}, \tag{A3}$$

and

$$G_{c2} = \frac{P - v_{c2}^2\tilde{\rho}_{11}}{v_{c2}^2\tilde{\rho}_{12} - Q}. \tag{A4}$$

In the above equations v_{c1} and v_{c2} refer to the frequency-dependent wave velocities of the fast wave and the slow wave.

Assuming an $e^{i\omega t}$ temporal variation, the linearized radial momentum equation for the liquid phase can be written as follows:

$$\omega^2\phi\rho_f U_r = \phi\frac{\partial p_f}{\partial r} + [\tilde{\alpha}(\omega) - 1]\omega^2\phi\rho_f(u_r - U_r). \tag{A5}$$

The above equation is integrated from the bubble radius ($r = a$) to infinity in order to find an equation of motion for the bubble, which reads

$$\begin{aligned}\phi\rho_f\omega^2(G_{c1}\Phi_{c1} + G_{c2}\Phi_{c2}) = &-\phi(p_{f\infty} - p_{fa}) \\ &+ \omega^2\phi\rho_f[\tilde{\alpha}(\omega) - 1][\Phi_{c1a}(1 \\ &- G_{c1}) + \Phi_{c2a}(1 - G_{c2})].\end{aligned} \tag{A6}$$

We seek solutions for the potentials outside the bubble in the form

$$\Phi_{c1} = \frac{A_{c1}e^{-ik_1 r}}{r}, \tag{A7}$$

and

$$\Phi_{c2} = \frac{A_{c2}e^{-ik_2 r}}{r}, \tag{A8}$$

where k_1 and k_2 are the radial wave numbers associated with the fast compressional wave and the slow compressional wave respectively. Then, substitution of the solutions given by Eqs. (A7) and (A8) into Eq. (A6), leads to the momentum equation in terms of the two unknowns A_{c1} and A_{c2}. The boundary conditions at the bubble surface provide the remaining relations to close the problem. Inside the bubble we neglect the interaction between the air and the solid matrix and the matrix is considered as acoustically compact. It can be shown that this condition implies that the velocity of the solid phase linearly depends on r. We assume continuity of the radial velocity of the solid phase and its radial derivative across the bubble surface. This last condition allows a closed analytical solution and it is consistent with numerical calculations based on the gas pocket model.[5,6] This leads to the following relation:

$$\frac{\partial^2 u_r}{\partial r\partial t}(a^+) = \frac{1}{r}\frac{\partial u_r}{\partial t}(a^+), \tag{A9}$$

which holds at the outside of the bubble (a^+). The continuity of fluid volume provides an equation for the change in the volume of the gas bubble ΔV_g in terms of the fluid and solid displacements at the bubble surface

$$\Delta V_g = 4\pi a^2[(1 - \phi)u_r + \phi U_r]. \tag{A10}$$

We also consider that the pressure difference across the bubble surface is balanced by the radial viscous stress in the fluid at the bubble surface

$$p_f(a^+) - p_g = \frac{4}{3}\eta\frac{\partial^2 U_r}{\partial r\partial t}(a), \tag{A11}$$

where $p_f(a^+)$ denotes the pressure outside the bubble evaluated at the bubble radius and p_g is the gas pressure inside the bubble.

Substitution of the expressions for Φ_{c1} and Φ_{c2} in the boundary conditions [Eqs. (A9)–(A11)], followed by some algebraic manipulations lead to the following relation between the volume of the air bubble and the external pressure $p_{f\infty}$:

$$\begin{aligned}&\omega^2\rho_f\left[\frac{a_1b_2 - a_2b_1}{a_1c_2 - a_2c_1} + \frac{4}{3}\frac{i\eta\phi a_1a_2(G_{c2} - G_{c1})}{\omega\rho_f(a_1c_2 - a_2c_1)}\right]\frac{V_g}{4\pi\phi a} \\ &\quad = p_{f\infty} - p_g,\end{aligned} \tag{A12}$$

where

$$a_j = k_j^2\left(1 - 3\frac{1 + ik_j a}{k_j^2 a^2}\right), \tag{A13}$$

$$b_j = \phi G_{cj} - (G_{cj} - 1)\frac{\tilde{\rho}_{12} + \phi\rho_f\tilde{\alpha}(\omega)}{\rho_f}, \tag{A14}$$

and

$$c_j = (1 + ik_j a)(1 - \phi + \phi G_{cj}). \tag{A15}$$

The last dissipative mechanism considered in this model is the thermal damping. It arises due to the heat exchange between the gas phase and the solid matrix induced by the oscillations of the bubble. Its contribution to the bulk modulus of the gas phase can be expressed as np_g. Here we have introduced a complex-valued polytropic coefficient, n:

$$\begin{aligned}n = \gamma\Bigg(1 + 3(\gamma - 1)\Bigg\{&\frac{\coth[\psi(8\alpha_\infty k_0/\phi)^{1/2}]}{\psi(8\alpha_\infty k_0/\phi)^{1/2}} \\ &- \frac{1}{[\psi(8\alpha_\infty k_0/\phi)^{1/2}]^2}\Bigg\}\Bigg)^{-1},\end{aligned} \tag{A16}$$

where $\psi = (1+i)(\omega/2a_g)$, a_g being the thermal diffusivity of the gas and γ being the specific heat ratio of the gas (for air $\gamma = 1.4$). Champoux and Allard[22] and Henry *et al.*[23] reported a slightly different expression for the polytropic exponent n.

Finally, the following expression is found for the frequency-dependent bulk modulus of the gas phase $K_g = -V_g(\partial V_g/\partial p_{f\infty})^{-1}$:

$$K_g(\omega) = \frac{1}{3}a^2\omega^2\rho_f\left[\frac{3np_g}{a^2\omega^2\rho_f} - \frac{a_1b_2 - a_2b_1}{a_1c_2 - a_2c_1} - \frac{4}{3}\frac{i\eta\phi a_1a_2(G_{c2} - G_{c1})}{\omega\rho_f(a_1c_2 - a_2c_1)}\right]. \tag{A17}$$

APPENDIX B: DISPLACEMENT POTENTIAL FORMULATION FOR THE SURFACE MODES

In this appendix a displacement potential formulation is developed in order to describe the surface waves that propagate along a plane interface between a fluid half-space and a liquid-saturated porous half-space. The configuration studied is displayed in Fig. 3.

The surface modes propagate parallel to the interface, depend exponentially on the distance z from the interface and can be expressed in terms of the bulk mode solutions. In the liquid ($z<0$), the compressional waves are described by the following potential:

$$\Phi_f = A_f e^{\gamma_f z} e^{i(k_x x - \omega t)}. \tag{B1}$$

The potentials associated to each of the bulk modes which propagate in the porous half-space are

$$\Phi_{c1} = A_{c1} e^{-\gamma_{c1} z} e^{i(k_x x - \omega t)}, \tag{B2}$$

$$\Phi_{c2} = A_{c2} e^{-\gamma_{c2} z} e^{i(k_x x - \omega t)}, \tag{B3}$$

and

$$\mathbf{\Psi_{sh}} = B e^{-\gamma_{sh} z} e^{i(k_x x - \omega t)} \hat{e}_y, \tag{B4}$$

where $\hat{e}_y$ is the cartesian basis vector in the y direction. The above potentials describe waves that propagate parallel to the interface. The wave numbers in the z direction are related to the horizontal wave number k_x through the following relations:

$$\gamma_j = \sqrt{k_x^2 - \frac{\omega^2}{c_j^2}}, \quad j = 1,2,\mathrm{sh},f, \tag{B5}$$

where c_j is the velocity of the corresponding bulk mode.

The surface modes can be written as a frequency-dependent linear combination of the potentials stated above. The different contributions of the bulk modes are determined by the boundary conditions, namely: continuity of averaged normal displacement, total stress, and pressure. The displacements of the solid phase and the fluid phase in the porous medium can be expressed as follows:

$$\boldsymbol{u} = \boldsymbol{\nabla}(\Phi_{c1} + \Phi_{c2}) + \boldsymbol{\nabla} \times \mathbf{\Psi_{sh}}, \tag{B6}$$

and

$$\boldsymbol{U} = G_{c1}\boldsymbol{\nabla}\Phi_{c1} + G_{c2}\boldsymbol{\nabla}\Phi_{c2} + G_{sh}\boldsymbol{\nabla} \times \mathbf{\Psi_{sh}}, \tag{B7}$$

where $\boldsymbol{u}$ refers to the displacement of the matrix and $\boldsymbol{U}$ to the displacement of the pore fluid. In the liquid half-space, the displacement $\boldsymbol{U}_f$ is $\boldsymbol{\nabla}\Phi_f$. Therefore the continuity of average normal displacement at the interface

$$(1-\phi)u_z + \phi U_z = U_{fz}, \tag{B8}$$

can be expressed as:

$$(1 - \phi + \phi G_{c1})\frac{\partial\Phi_{c1}}{\partial z} + (1 - \phi + \phi G_{c2})\frac{\partial\Phi_{c2}}{\partial z} + (1 - \phi + \phi G_{sh})\frac{\partial\Psi_{sh}}{\partial x} = \frac{\partial\Phi_f}{\partial z}. \tag{B9}$$

The continuity of the normal component of the total stress implies

$$\tau_{zz} - \phi p = -p_f. \tag{B10}$$

Using the Biot's stress-strain relations (see, e.g., Allard[19]), the above equation can be written in terms of the potentials as follows:

$$[P - 2N + Q + G_{c1}(Q+R)]\boldsymbol{\nabla}^2\Phi_{c1} + 2N\frac{\partial^2\Phi_{c1}}{\partial z^2} + [P - 2N + Q + G_{c2}(Q+R)]\boldsymbol{\nabla}^2\Phi_{c2} + 2N\frac{\partial^2\Phi_{c2}}{\partial z^2} + 2N\frac{\partial^2\Psi_{sh}}{\partial z\partial x} = -\omega^2\rho_w\Phi_f. \tag{B11}$$

The absence of tangential stress in the liquid requires $\tau_{xz} = 0$ at the interface, and this condition implies that

$$N\left[2\left(\frac{\partial^2\Phi_{c1}}{\partial z\partial x} + \frac{\partial^2\Phi_{c2}}{\partial z\partial x}\right) + \frac{\partial^2\Psi_{sh}}{\partial x^2} - \frac{\partial^2\Psi_{sh}}{\partial z^2}\right] = 0. \tag{B12}$$

Finally, the continuity of pressure leads to

$$-\frac{1}{\phi}[(Q + RG_{c1})\boldsymbol{\nabla}^2\Phi_{c1} + (Q + RG_{c2})\boldsymbol{\nabla}^2\Phi_{c2}] = \rho_w\omega^2\Phi_f. \tag{B13}$$

Substituting Eqs. (B1)–(B4) into Eqs. (B9) and (B11)–(B13) and after some algebraic manipulations a linear system for the amplitudes of the potentials is found

$$\mathbf{N}(k_x,\omega)\cdot\mathbf{a} = 0, \tag{B14}$$

where the matrix $\mathbf{N}$ contains information about the mechanical properties of the fully saturated porous medium and the water half-space and $\mathbf{a}$ is a vector containing the amplitude of the wave potentials, $\mathbf{a}^T = (A_f, A_{c1}, A_{c2}, B)$. The elements of the matrix $\mathbf{N}$ are given in Appendix C. The surface modes satisfy the condition that the determinant of $\mathbf{N}$ equals zero

$$\det[\mathbf{N}(k_x,\omega)] = 0. \tag{B15}$$

At a fixed frequency ω, Eq. (B15) is numerically solved for complex k_x using a Newton-Raphson algorithm. In this way, frequency-dependent phase velocities, $V(\omega) = \omega\,\mathrm{Re}^{-1}(k_x)$, and specific attenuation coefficients, $Q^{-1}(\omega) = |2\,\mathrm{Im}(k_x)/\mathrm{Re}(k_x)|$ are obtained. The frequency-dependent bulk modulus of the mixture of fluids $K_f(\omega)$, enters the model via the generalized elastic coefficients P, Q, and R, which themselves also influence the complex velocities of the compressional waves $c1$ and $c2$.

APPENDIX C: MATRIX COEFFICIENTS

In this appendix the elements n_{ij} of the matrix **N** of Eq. (B14) are explicitly given

$$n_{11} = \gamma_f,$$

$$n_{12} = \gamma_{c1}(1 - \phi + \phi G_{c1}),$$

$$n_{13} = \gamma_{c2}(1 - \phi + \phi G_{c2}),$$

$$n_{14} = -ik_x(1 - \phi + \phi G_{\mathrm{sh}}),$$

$$n_{21} = 0,$$

$$n_{22} = 2N\gamma_{c1}ik_x,$$

$$n_{23} = 2N\gamma_{c2}ik_x,$$

$$n_{24} = (\gamma_{\mathrm{sh}}^2 + k_x^2)N,$$

$$n_{31} = \omega^2\rho_f,$$

$$n_{32} = -[(P - 2N) + Q + G_{c1}(Q + R)]\left(\frac{\omega}{c_1}\right)^2 + 2N\gamma_{c1}^2,$$

$$n_{33} = -[(P - 2N) + Q + G_{c2}(Q + R)]\left(\frac{\omega}{c_2}\right)^2 + 2N\gamma_{c2}^2,$$

$$n_{34} = -2Nik_x\gamma_{\mathrm{sh}},$$

$$n_{41} = \omega^2\rho_f,$$

$$n_{42} = -\left(\frac{\omega}{c_1}\right)^2\frac{1}{\phi}(Q + RG_{c1}),$$

$$n_{43} = -\left(\frac{\omega}{c_2}\right)^2\frac{1}{\phi}(Q + RG_{c2}),$$

$$n_{44} = 0.$$

[1]L. Van Wijngaarden, "One dimensional flow of liquids containing small gas bubbles," Annu. Rev. Fluid Mech. **4**, 369–395 (1972).
[2]M. A. Biot, "Theory of propagation of elastic waves in a fluid-saturated porous solid. I: Low-frequency range," J. Acoust. Soc. Am. **28**, 168–178 (1956).
[3]M. A. Biot, "Theory of propagation of elastic waves in a fluid-saturated porous solid. II: Higher frequency range," J. Acoust. Soc. Am. **28**, 179–191 (1956).
[4]J. E. White, "Computed seismic speeds and attenuation in rocks with partial gas saturation," Geophysics **40**, 224–232 (1975).
[5]N. C. Dutta and H. Ode, "Attenuation and dispersion of compressional waves in fluid-filled porous rocks with partial gas saturation (White model)—Part I: Biot theory," Geophysics **44**, 1777–1788 (1979).
[6]N. C. Dutta and H. Ode, "Attenuation and dispersion of compressional waves in fluid-filled porous rocks with partial gas saturation (White model)—Part II: Results," Geophysics **44**, 1789–1805 (1979).
[7]J. G. Berryman, L. Thigpen, and R. C. Y. Chin, "Bulk elastic wave propagation in partially saturated porous solids," J. Acoust. Soc. Am. **84**, 360–373 (1988).
[8]D. M. J. Smeulders and M. E. H. van Dongen, "Wave propagation in porous media containing a dilute gas-liquid mixture: Theory and experiments," J. Fluid Mech. **343**, 351–373 (1997).
[9]D. L. Johnson, "Theory of frequency dependent acoustics in patchy-saturated porous media," J. Acoust. Soc. Am. **110**, 682–694 (2001).
[10]J. M. Carcione, H. B. Helle, and N. H. Pham, "White's model for wave propagation in partially saturated rocks: Comparison with poroelastic numerical experiments," Geophysics **68**, 1389–1398 (2003).
[11]G. Degrande, G. De Roeck, P. Van Den Broeck, and D. Smeulders, "Wave propagation in layered dry, saturated and unsaturated poroelastic media," Int. J. Solids Struct. **35**, 4753–4778 (1998).
[12]G. A. Gist, "Interpreting laboratory velocity measurements in partially gas-saturated rocks," Geophysics **59**, 1100–1109 (1994).
[13]T. Cadoret, D. Marion, and B. Zinszner, "Influence of frequency and fluid distribution on elastic wave velocities in partially saturated limestones," J. Geophys. Res., [Solid Earth] **100**, 9789–9803 (1995).
[14]T. Cadoret, G. Mavko, and B. Zinszner, "Fluid distribution effect on sonic attenuation in partially saturated limestones," Geophysics **63**, 154–160 (1998).
[15]N. Lucet, P. N. Rasolofosaon, and B. Zinszner, "Sonic properties of rocks under confining pressure using the resonant bar technique," J. Acoust. Soc. Am. **89**, 980–990 (1991).
[16]S. Feng and D. L. Johnson, "High-frequency acoustic properties of a fluid/porous solid interface. I: New surface mode," J. Acoust. Soc. Am. **74**, 906–914 (1983).
[17]S. Feng and D. L. Johnson, "High-frequency acoustic properties of a fluid/porous solid interface. II: The 2D reflection green's function," J. Acoust. Soc. Am. **74**, 915–914 (1983).
[18]A. A. Gubaidullin, O. Yu. Kuchugurina, D. M. J. Smeulders, and C. J. Wisse, "Frequency-dependent acoustic properties of a fluid/porous solid interface," J. Acoust. Soc. Am. **116**, 1474–1980 (2004).
[19]J. F. Allard, *Propagation of Sound in Porous Media* (Elsevier Science, New York, 1993).
[20]Th. Bourbié, O. Coussy, and B. Zinszner, *Acoustics of Porous Media* (Gulf, Houston, 1987).
[21]A. B. Wood, *A Textbook of Sound* (Bell, London, 1955).
[22]Y. Champoux and J. F. Allard, "Dynamic tourtuosity and bulk modulus in air-saturated porous media," J. Appl. Phys. **70**, 1975–1797 (1991).
[23]M. Henry, P. Lemarinier, J. F. Allard, J. L. Bonardet, and A. Gedeon, "Evaluation of the characteristic dimensions for porous sound-absorbing materials," J. Appl. Phys. **77**, 17–20 (1995).

Resonance modes in a one-dimensional medium with two purely resistive boundaries: Calculation methods, orthogonality, and completeness

Jean Kergomard[a] and Vincent Debut
Laboratoire de Mécanique et d'Acoustique, CNRS UPR 7051, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

Denis Matignon
Télécom Paris/Département TSI, CNRS UMR 5141 37-39, rue Dareau 75014 Paris, France

(Received 25 April 2005; revised 22 December 2005; accepted 23 December 2005)

Studying the problem of wave propagation in media with resistive boundaries can be made by searching for "resonance modes" or free oscillations regimes. In the present article, a simple case is investigated, which allows one to enlighten the respective interest of different, classical methods, some of them being rather delicate. This case is the one-dimensional propagation in a homogeneous medium having two purely resistive terminations, the calculation of the Green function being done without any approximation using three methods. The first one is the straightforward use of the closed-form solution in the frequency domain and the residue calculus. Then, the method of separation of variables (space and time) leads to a solution depending on the initial conditions. The question of the orthogonality and completeness of the complex-valued resonance modes is investigated, leading to the expression of a particular scalar product. The last method is the expansion in biorthogonal modes in the frequency domain, the modes having eigenfrequencies depending on the frequency. Results of the three methods generalize or/and correct some results already existing in the literature, and exhibit the particular difficulty of the treatment of the constant mode. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2166709]

PACS number(s): 43.20.Ks, 43.40.Cw, 02.30.Tb, 02.30.Jr [MO] Pages: 1356–1367

I. INTRODUCTION

Studying the problem of wave propagation in media with resistive boundaries can be made by searching for "resonance modes" (see Ref. 1), or free oscillations regimes. These modes can be nonorthogonal for the ordinary scalar product, entailing some difficulties depending on the mathematical treatment, made either in the time or frequency domain. Two classical methods exist for such a problem, and can be used either for a scalar, second-order differential equation, or for a system of two equations of the first order. They have been especially used for the problem of a one-dimensional (1D) medium with one resistive boundary, the other boundary condition being of Dirichlet type:

(i) In the time domain, the use of time and space as separate variables leads directly to the basis of modes, but they are nonorthogonal for the most common product, and difficulties occur when searching for the coefficients depending, for instance, on initial conditions. Nevertheless, for a particular case, Oliverto and Santini,[2] and Guyader[3] have solved the problem, and Rideau,[4] using a system of equations of the first order, found a scalar product making the modes orthogonal (see also Refs. 5–7), and gave the proof of completeness.

(ii) In the frequency domain, the equations to be solved are ordinary differential equations with boundary conditions depending on frequency, but the use of orthogonal decomposition is possible. This leads to eigenmodes and eigenfrequencies *depending on frequency*. It is the case for the classical theory of room acoustics (see, e.g., Morse and Ingard),[8] using biorthogonality. To return to time domain in order to deduce the resonance modes is a rather delicate task, especially because of the calculation of the derivation of eigenfrequencies with respect to frequency. Biorthogonality has been used also for duct modes (see, e.g., Ref. 9) Another approach has been recently used by Trautmann and Rabenstein,[10,11] using a system of first-order equations (these authors treat the case of two resistive boundary conditions).

The present article is devoted to the study of the simple 1D case, when the two boundaries are resistive. One goal is to exhibit how the different methods articulate. We start by using the fact that a straightforward solution exists for the wave equation with source, by applying the residue calculus to the closed form of the Fourier domain solution: As discussed by Levine,[12] this closed-form solution, avoiding the sum of a series, is "relatively poorly, if not entirely, unknown to the general acoustics community." All calculations can be carried out analytically without any approximation, exhibiting the properties of the different methods (however many previous papers restrict their content to small impedance, or admittance, at one extremity, using perturbation methods). The case under study corresponds to 1D propagation in a homogeneous medium bounded by two other semi-infinite media with different characteristic impedances, dissipation being therefore due to radiation at infinity. It is especially

[a]Electronic address: kergomard@lma.cnrs-mrs.fr

 0001-4966/2006/119(3)/1356/12/$22.50

interesting because of its physical significance (it is probably the simplest radiation problem), and also because it realizes one of the possible transitions between Neumann and Dirichlet boundary conditions. Notice that in the context of optics and quantum mechanics, the problem has been studied including the outside media by Leung *et al.*[13,14] the resonance modes being called quasinormal modes.

In Sec. II, the equations to be solved are stated, with some possible physical interpretations. As a first step, the classical, closed-form solution of the Green function in the frequency domain is established (Sec. III), with its inverse Fourier transform (FT) corresponding to the successive reflections (Sec. IV). The second step is the residue calculus in order to determine the resonance modes [Sec. V, the basic result being given by Eq. (26)]. Then, results are compared to those of the two aforementioned methods, i.e.:(i) The method of separation of variables (Sec. VI), which gives the result for given initial conditions [the corresponding results being Eqs. (37), (48), and (49)]; in this section, the question of orthogonality and completeness of the modes is investigated; and (ii) the method of eigenmodes in the frequency domain (section VII). For the two methods, both second-order scalar equation and first-order system of two equations are used successively, with emphasis on the existence of a constant mode.

II. STATEMENT OF THE PROBLEM, PHYSICAL INTERPRETATION

The Green function $g(x,t|x_0,t_0)$ for the wave equation is solution of the following equation:

$$[\partial_{xx}^2 - c^{-2}\partial_{tt}^2]g(x,t) = -\delta(x-x_0)\delta(t-t_0), \tag{1}$$

where x and x_0 are the spatial coordinates of the receiver and source, respectively (or vice versa), t and t_0 the times of observation and excitation, respectively, c is the speed of sound. $\delta(x)$ is the Dirac function.

For sake of simplicity, x_0 and t_0 are considered to be fixed. Moreover in the whole paper, the choice of $t_0=0$ is made. For negative t, the function is zero, as well as its first derivative. The Green function satisfies the following boundary conditions:

$$c\zeta\partial_x g(x,t) = \partial_t g(x,t) \text{ at } x=0, \tag{2}$$

$$c\zeta_\ell\partial_x g(x,t) = -\partial_t g(x,t) \text{ at } x=\ell, \tag{3}$$

where $\zeta=Z/\rho c$, ρ is the density of the fluid, and Z is the impedance at $x=0$, which is assumed to be a real quantity, independent of the frequency. Similarly, $\zeta_\ell=Z_\ell/\rho c$, where Z_ℓ is the impedance at $x=\ell$ (ℓ being positive).

An obvious physical interpretation for quantities ζ and ζ_ℓ is the following: Consider for $x<0$ and $x>\ell$ (see Fig. 1) two media with characteristic impedances $\rho_- c_-$ and $\rho_+ c_+$, respectively. If the media are nondissipative, impedances are real, and can be larger or smaller than the impedance of the bounded medium, ρc. Moreover, they are positive, because they correspond to waves outgoing from the bounded medium. Therefore, this is the problem of planar pressure waves in a stratified medium, the direction of propagation being normal to the interfaces. A generalization to more complex stratified media would be possible, at least numerically. In this problem, the Green function corresponds to the acoustic pressure: Of course, it does not have the dimension of a pressure, but the solution for a "concrete" problem with source can be easily solved, as explained in standard textbooks, and discussed in a recent paper by Levine.[12]

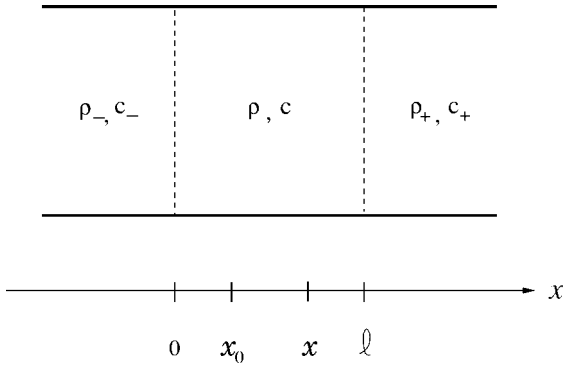


FIG. 1. 1D medium bounded with two other media.

Other problems correspond to the previous equations:

(i) In an approximate way, ignoring higher-order duct modes, the problem of planar waves in a rigid walled duct terminating in two semi-infinite ducts with different cross sections areas, the quantities ζ and ζ_ℓ being the ratios of the areas. The approximation is good at low frequencies.

(ii) The problem of a dissipative termination: The terminal impedances Z and Z_ℓ can be the impedances of dissipative media (at low frequencies, a porous medium open to a large space can be an approximation of a pure resistance, due to viscous effects).

In all of the previous problems, the quantities ζ and ζ_ℓ are real and positive, the terminations being passive. For active terminations, they can be negative. An example is the beginning of self-sustained oscillations in musical instruments: A nonlinear excitator, such as a reed for a clarinet, can be linearized as a pure resistance. When the main control parameter, i.e., the pressure in the mouth of the musician, increases, the resistance becomes negative, the static regime becomes unstable, and an oscillation starts as an increasing exponential (see e.g., Refs. 15 and 16).

Obviously, analogous problems for mechanical vibrations or other wave fields are numerous.

III. CLOSED-FORM SOLUTION FOR THE FOURIER TRANSFORM

The FT of $g(x,t)$ is denoted $G(x,\omega)$ (throughout the article, functions of time are written in small characters, and their FTs are written in capital characters). It is equal to

$$G(x,\omega) = \int_{-\infty}^{+\infty} g(x,t)e^{-i\omega t}dt, \quad \text{where} \tag{4}$$

$$g(x,t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} G(x,\omega)e^{i\omega t}d\omega. \tag{5}$$

The FT of Eqs. (1)–(3) are found to be

$$(\partial_{xx}^2 + \omega^2/c^2)G(x,\omega) = -\delta(x-x_0), \tag{6}$$

$$c\zeta\partial_x G(x,\omega) = i\omega G(x,\omega) \text{ at } x=0; \tag{7}$$

$$c\zeta_\ell\partial_x G(x,\omega) = -i\omega G(x,\omega) \text{ at } x=\ell. \tag{8}$$

While terminal impedances are independent of frequency; boundary conditions are frequency dependent. Nevertheless a classical closed-form solution is already known, which has been especially used in Ref. 17. If $x \neq x_0$ solutions of Eq. (6) can be written as

$$G(x,\omega) = A^- \cosh[i\omega x/c + \eta], \quad \text{if } x < x_0;$$

$$G(x,\omega) = A^+ \cosh[i\omega(\ell - x)/c + \eta_\ell], \quad \text{if } x > x_0.$$

For the boundary conditions, the following definitions are used:

$$\zeta = \coth\ \eta; \quad r = e^{-2\eta} = (\zeta - 1)/(\zeta + 1),$$

$$\zeta_\ell = \coth\ \eta_\ell; \quad r_\ell = e^{-2\eta_\ell} = (\zeta_\ell - 1)/(\zeta_\ell + 1), \tag{9}$$

where r and r_ℓ are the reflection coefficients. The quantity η satisfies: $2\eta = -(\ln|r| + i\arg(r))$ $[2\pi]$. Because r is real, we choose the following definition:

$$\eta = \eta_r + i\mu\pi/2; \quad \mu = 0 \text{ or } 1. \tag{10}$$

Two cases exist: (i) if $|\zeta| > 1$, $r > 0$, $\mu = 0$; (ii) if $|\zeta| < 1$, $r < 0$, $\mu = 1$. A similar remark and definition can be applied to boundary $x = \ell$:

$$\eta_\ell = \eta_{\ell r} + i\mu_\ell\pi/2; \quad \mu_\ell = 0 \text{ or } 1. \tag{11}$$

The case $\zeta = 1$ (semi-infinite tube or medium) corresponds to $\eta = \infty$: It is discussed in the next sections. Except for the last one, most of the following calculations are valid for all cases. At $x = x_0$, writing the continuity of the function and the jump of its first derivative, the following result is obtained:

$$G(x,\omega) = \frac{c}{i\omega}\frac{\cosh[\eta + i\omega x_0/c]\cosh[\eta_\ell + i\omega(\ell - x)/c]}{\sinh(i\omega\ell/c + \eta + \eta_\ell)}, \tag{12}$$

if $x \geqslant x_0$ and a similar result if $x \leqslant x_0$, by interchanging x and x_0.

IV. SOLUTION IN THE TIME DOMAIN (SUCCESSIVE REFLECTIONS)

Equation (12) can be transformed in the time domain, leading to a solution corresponding to the successive reflections of the Green function in infinite space at the two boundaries. It will be the reference solution for the check of the validity of the modal expansion. The sinh function of the denominator can be written as

$$\sinh(i\omega\ell/c + \eta + \eta_\ell) = \frac{1 - e^{-2\eta - 2\eta_\ell - 2i\omega\ell/c}}{2e^{-\eta - \eta_\ell - i\omega\ell/c}},$$

and, if the modulus of the exponential at the denominator is less than unity (this is discussed hereafter), as

$$\sinh^{-1}(i\omega\ell/c + \eta + \eta_\ell) = 2e^{-\eta - \eta_\ell - i\omega\ell/c}[1 + F(\omega) + F^2(\omega) + F^3(\omega) + \cdots]. \tag{13}$$

$F(\omega) = \exp(-2\eta - 2\eta_\ell - 2i\omega\ell/c)$ is the function corresponding to a complete round trip of a wave in the tube, of duration $2\ell/c$. Concerning the numerator of Eq. (12), it can be written: $\exp(+\eta + \eta_\ell + i\omega\ell/c)G_p(x,\omega)c/4$, where:

$$G_p(x,\omega) = e^{-i\omega(x-x_0)/c} + re^{-i\omega(x+x_0)/c} + r_\ell e^{-i\omega(2\ell - x - x_0)/c} + rr_\ell e^{-i\omega(2\ell - x + x_0)/c}. \tag{14}$$

Therefore, the Green function is:

$$G(x,\omega) = \frac{c}{2i\omega}G_p(x,\omega)[1 + F(\omega) + F^2(\omega) + \ldots]. \tag{15}$$

The factor $G_p(x,\omega)/i\omega$ corresponds to the four "primary" waves arriving during the first cycle of duration $2\ell/c$, and this packet is simply reproduced at times $2\ell/c$, $4\ell/c$, $6\ell/c$, etc. (for a detailed explanation, see e.g., Kergomard).[15] The inverse FT of the function $G_p(x,\omega)/i\omega$, denoted $h_p(x,t)$, is obtained by taking into account the zero condition for negative times. The result is found to be, whatever the sign of $(x - x_0)$:

$$h_p(x,t) = H[t - |x - x_0|/c] + rH[t - (x + x_0)/c] + rH[t - (2\ell - x - x_0)/c] + rr_\ell H[t - (2\ell - |x - x_0|)/c], \tag{16}$$

where $H(t)$ is the step function. Finally

$$g(x,t) = \frac{c}{2}h_p(x,t) * [\delta(t) + f(t) + (f * f)(t) + \ldots]; \tag{17}$$

$$f(t) = rr_\ell\delta(t - 2\ell/c). \tag{18}$$

The condition of validity of expansion (13) is $|rr_\ell| < 1$. We notice that if ζ is real and positive, $|r| < 1$, and similarly for ζ_ℓ. Therefore, the condition is satisfied when the two boundaries are dissipative, or, more precisely, if the combination of the two reflections is dissipative. The case $|rr_\ell| > 1$ will be discussed in Sec. V C. Other comments can be made:

(a) The article is limited to purely resistive boundaries, but Eqs. (16) and (18) can be generalized to various boundary conditions defined by a reflection coefficient, $R(\omega)$. This is done by replacing the products like $rH(t)$ by the convolution product $(r * H)(t)$, where $r(t)$ is the inverse FT of $R(\omega)$.

(b) For the case under study, we notice that the convolution product of n times function $f(t)$ is $(rr_\ell)^n\delta(t - 2n\ell/c)$.

(c) If ζ (respectively ζ_ℓ) is unity, the reflection coefficient r (respectively r_ℓ) vanishes, as well as $f(t)$: The first term of the Green function is the Green function of an infinite medium, the first two terms correspond to a semi-infinite medium, etc… As it will be seen in the next section, no modes can be found for these cases, because no reflections exist, either η or η_ℓ tending to infinity.

(d) Finally, multiplying both members of Eq. (15) by the factor $[1 - F(\omega)]$, and taking the inverse FT, it appears

that a closed-form exists in the time domain, which is the basis for the study of the Helmholtz motion of bowed string instruments (see, e.g., Woodhouse).[18] It is a recurrence relationship:

$$\partial_t g(x,t) - rr_\ell \partial_t g(x,t-2\ell/c) = g_p(x,t)c/2.$$

V. EXPANSION IN RESONANCE MODES USING THE INVERSE FT

Putting expression (12) of the frequency domain in Eq. (5) leads to the modal expansion of the time domain expression. The tool is the residue calculus. If all poles of expression (12) are simple and located on or above the real axis, the following equation can be used:

$$g(x,t) = i\Sigma, \quad \text{if } t>0 \text{ and } 0 \text{ if } t<0, \tag{19}$$

where Σ is the sum of the residues of $G(x,\omega)\exp(i\omega t)$ (see e.g., Morse and Ingard[8] p. 17, changing i to $-i$).

A. Calculation of the poles

Zeros of function sinh satisfy:

$$i\omega_n = [-\eta - \eta_\ell + in\pi]c/\ell, \tag{20}$$

where n is an integer. In order for the poles to be above the real axis, the condition is $\eta_r + \eta_{\ell r} > 0$. It is equivalent to the condition previously obtained for the successive reflections expansion: $|rr_\ell| < 1$. Using definition (10), Eq. (20) is rewritten as

$$\omega_n = [n - (\mu + \mu_\ell)/2]\pi c/\ell + i(\eta_r + \eta_{\ell r})c/\ell. \tag{21}$$

As already remarked by several authors, the imaginary part of the complex frequency is independent of n, and the real part is independent of the dissipation. Depending on the values of ζ and ζ_ℓ, different cases must be distinguished:

(i) If $|\zeta|>1$ and $|\zeta_\ell|>1$ (real η and η_ℓ): The real part of the frequency corresponds to the values for pure Neumann conditions (infinite ζ and ζ_ℓ).

(ii) If $|\zeta|>1$ and $|\zeta_\ell|<1$ (mixed case with either complex η or complex η_ℓ: either μ or μ_ℓ is unity): The real part corresponds to a problem with different conditions (Neumann and Dirichlet) at $x=0$ and $x=\ell$. The real part of eigenfrequencies is an odd harmonic of $c/4\ell$.

(iii) If $|\zeta|<1$ and $|\zeta_\ell|<1$ (complex η and η_ℓ: $\mu=\mu_\ell=1$): the real part corresponds to the values for pure Dirichlet conditions (zero ζ and ζ_ℓ).

Except for case (ii), a purely imaginary eigenfrequency exists for $n=(\mu+\mu_\ell)/2$.

B. Calculation of the residues

In all cases, the Taylor expansion of the function sinh in Eq. (12) at the first order of the quantity $(\omega-\omega_n)$ can be determined. The result is

$$\sinh[i\omega\ell/c + \eta + \eta_\ell] \simeq i(-1)^n(\omega-\omega_n)\ell/c. \tag{22}$$

We get for ω close to the pole ω_n:

$$G(x,\omega) = -\frac{c^2}{\omega_n \ell}\frac{f_n(x)f_n(x_0)}{(\omega-\omega_n)}; \tag{23}$$

$$f_n(x) = \cosh(\eta + i\omega_n x/c), \tag{24}$$

or $f_n(x)=(-1)^n\cosh(i\omega_n(\ell-x)/c+\eta_\ell)$. The residue corresponding to the pole $\omega=0$, remains to be calculated. For small ω,

$$G(x,\omega) = \frac{c}{i\omega}\frac{\cosh\eta\cosh\eta_\ell}{\sinh(\eta+\eta_\ell)} = \frac{c}{i\omega}\frac{1}{\zeta^{-1}+\zeta_\ell^{-1}}. \tag{25}$$

Using Eq. (19), the inverse FT of $G(x,\omega)$ is obtained:

$$g(x,t) = H(t)\frac{c^2}{\ell}\sum_n \frac{f_n(x)f_n(x_0)}{i\omega_n}e^{i\omega_n t} + \frac{cH(t)}{\zeta^{-1}+\zeta_\ell^{-1}}. \tag{26}$$

Some comments can be made.

(a) The formula is valid for all aforementioned cases.

(b) The mode shapes $f_n(x)$ are complex-valued functions of the space variable, meaning that the shape is varying with time (for a discussion on complex modes, see e.g; Ref. 19). The question of their orthogonality will be discussed in Sec. VI. Notice that functions $f_n(x)$ do not fulfill the same boundary conditions than $G(x,\omega)$: the boundary conditions are (7) and (8), but where ω is replaced by ω_n.

(c) The last term in Eq. (26) is a constant mode. If one of the impedances ζ or ζ_ℓ is zero, it disappears, as it is intuitive, in order to satisfy a Dirichlet condition. It is a trivial solution of the wave equation and the boundary conditions, and can be compared to the dc component of a periodic signal. When both ζ and ζ_ℓ tend to infinity, the boundaries tend to Neumann boundaries, and the combination of the nonoscillatory mode of frequency ω_0 and the constant mode results in a uniform (i.e., constant in space) mode which increases linearly with time. The calculation is done as follows: If η and η_ℓ tend to zero, ω_0 tends to zero, and the factor $\exp(i\omega_0 t)$ can be written as: $1+i\omega_0 t$. The zeroth-order term is equal to the opposite of the constant mode, and only the linear term remains. The result is $H(t)tc^2/\ell$, and its FT is $-c^2/\ell\omega^2$. This mode is the classical uniform mode existing for instance in three-dimensional cavities with rigid walls: Curiously, it exists in the standard textbooks (see, e.g., Ref. 8, page 571), but the time domain expression is not given. This mode is similar to the well-known planar guided mode, existing in ducts with rigid walls, whatever the geometrical shape.

(d) The imaginary part of the complex frequencies being independent of n, the decay is identical for all nonconstant modes.

(e) For the above-considered case (i), we notice that $\omega_{-n}=-\omega_n^*$ and $f_{-n}(x)=f_n^*(x)$, and, more generally,

$$i\omega_\nu = (i\omega_n)^*; \quad f_\nu(x) = (-1)^\mu f_n^*(x), \tag{27}$$

$$\text{where } \nu = -n + \mu + \mu_\ell \text{ is an integer.} \tag{28}$$

As a consequence, the solution $g(x,t)$ is real. It could be possible to transform the sum by adding the two oscillating terms corresponding to n and ν, when $n\neq\nu$, as it is usually done for nondissipative boundaries. Nevertheless, it appears that the formulas become intricate.

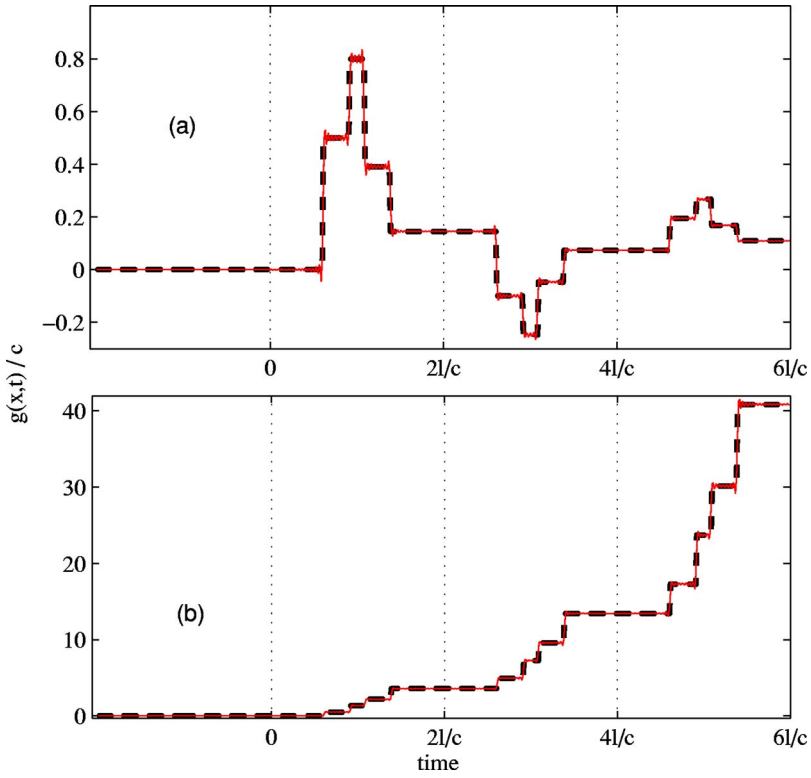


FIG. 2. (Color online) Normalized Green function as a function of time: comparison between the successive reflections method (dotted line) and modal expansion (solid line, 102 modes, i.e., maximum $n=50$). Locations of the source and receiver are $x_0/\ell=0.15$ and $x/\ell=0.76$, respectively. (a) Passive boundaries $\zeta=4$; $\zeta_\ell=0.1$ (constant mode=0.097). (b) Active boundary $\zeta=\zeta_\ell=-4$. (constant mode=−2).

(f) Equations (23) and (25) lead directly to another form of the FT of result (26), written as a series:

$$G(x,\omega) = -\frac{c^2}{\ell}\sum_n \frac{f_n(x)f_n(x_0)}{\omega_n(\omega-\omega_n)} + \frac{c}{i\omega}\frac{1}{\zeta^{-1}+\zeta_\ell^{-1}}. \tag{29}$$

An example of comparison of the successive reflections method and modal expansion is shown in Fig. 2(a), for conditions close to Neumann and Dirichlet. We notice that it is satisfactory. The Gibbs phenomenon appears, because of the truncated series of modes, ensuring the correct accordance between the two methods. Moreover, this accordance confirms the existence of the constant mode. Decreasing of the maxima is due to the dissipation at the boundaries: For pure Neumann and Dirichlet conditions, the shape would be similar, but perfectly periodical.

C. The case of active boundaries

What happens when the combination of boundaries is active, i.e., when $|rr_\ell|>1$, or $\eta_r+\eta_{\ell r}<0$ (at least one of the impedances ζ or ζ_ℓ is negative)? It is possible to prove that Eqs. (17) and (26) *remain valid for active boundary conditions*, as explained hereafter. The real part of $i\omega_n$ being independent of n, this suggests to use a new function $\tilde{g}(x,t)=g(x,t)\exp(-\tilde{\eta}t)$, where $\tilde{\eta}>-\eta_r-\eta_{\ell r}>0$, which can be substituted in the initial problem in order for the poles to be located again on or above the real axis. Equation (1) becomes

$$\partial_{xx}^2\tilde{g}(x,t) - c^{-2}[\partial_t+\tilde{\eta}]^2\tilde{g}(x,t) = -\delta(x-x_0)e^{-\tilde{\eta}t}\delta(t),$$

and similarly for Eqs. (2) and (3). It is equivalent to use an appropriate Laplace transform. Going in the frequency domain leads to Eqs. (6)–(8), where $G(x,\omega)$ is replaced by $\tilde{G}(x,\omega)$ and $i\omega$ by $(i\omega+\tilde{\eta})$, and a similar result for Eq. (12). The analysis of both successive reflections and poles and residues leads to the result $\tilde{g}(x,t)=g(x,t)\exp(-\tilde{\eta}t)$, where $g(x,t)$ is given by Eqs. (17) and (26), respectively, and the proof is achieved. Here, we do not repeat the complete procedure. We notice that for the case $\eta_r+\eta_{\ell r}=0$, one boundary is active and the other one is passive: Eigenfrequencies ω_n are real, while modes are complex. Figure 2(b) shows an example of result.

VI. METHOD OF SEPARATION OF VARIABLES

A. Second-order homogeneous equation with initial conditions

1. Derivation of the modes

Oliverto and Santini,[2] and Guyader[3] have treated a particular case of the problem (zero ζ, large ζ_ℓ) using the method of separation of variables. He gets nonorthogonal modes for the common scalar product $\int_0^\ell f_n(x)f_m(x)dx$. We will see that the method is valid whatever the values of the two boundary conditions, and how the derivation can be simplified.

We are searching for solutions $p(x,t)$ of homogeneous Eq. (1) (without second member), with boundary conditions (2) and (3), and with given initial conditions. Assuming that, the general solution is a superposition of solutions with separate variables, the solutions with separate variables are written in the following form:

$$p(x,t) = f(x)h(t); \tag{30}$$

$$h(t) = B^+e^{i\omega t} + B^-e^{-i\omega t}; \tag{31}$$

$$f(x) = \cosh(i\omega x/c + \varphi). \tag{32}$$

Decomposition (30) differs from the ordinary FT, because a priori ω is a complex quantity, depending on the boundary conditions. Considering first the solution $B^+e^{i\omega t}$, this leads to:

$$\zeta\omega \sinh \varphi = \omega \cosh \varphi; \tag{33}$$

$$\omega\zeta_\ell \sinh(i\omega\ell/c + \varphi) = -\omega \cosh(i\omega\ell/c + \varphi). \tag{34}$$

$\omega=0$ is a solution, corresponding to the constant mode. The other modes are given by Eq. (33): $\sinh(\varphi-\eta)=0$, thus:

$$f(x) = \cosh(i\omega x/c + \eta). \tag{35}$$

Actually, there is a sign $\pm$ in the right-hand side member of Eq. (35), but it is without importance, because it can be included in the coefficient B^+ of the solution. The eigenvalues equation is deduced from Eqs. (33) and (34), as follows:

$$\sinh(i\omega\ell/c + \eta + \eta_\ell) = 0, \tag{36}$$

the solutions being given by Eq. (20). The solution in time $B^-e^{-i\omega t}$ does not lead to new solutions for $f(x)$, therefore, assuming the solutions form a basis of solutions (this is discussed in Sec. VI B), the general solution of a problem with initial conditions can be written as:

$$p(x,t) = \sum_n A_n f_n(x) e^{i\omega_n t} + A, \tag{37}$$

where ω_n and $f_n(x)$ are given by Eqs. (20) and (24), respectively, and the coefficients A_n and A depend on the initial conditions, and can be determined using the orthogonality relation of the modes. A is the coefficient of the constant mode.

2. Orthogonality relationship between the modes: First approach

In order to derive an orthogonality relationship between the modes the common product is first calculated:

$$\Lambda_{nm} = \int_0^\ell f_n(x) f_m(x) dx. \tag{38}$$

Because $f_n(x)=(-1)^\mu f_\nu^*(x)$, the calculation of the quantities defined in Eq. (38) for all values of the index n is equivalent to the calculation of the quantities defined when replacing $f_m(x)$ by its conjugate. Writing

$$\int_0^\ell \left[f_n(x)\frac{d^2 f_m(x)}{dx^2} - f_m(x)\frac{d^2 f_n(x)}{dx^2} \right] dx = \left[f_n(x)\frac{df_m(x)}{dx} - f_m(x)\frac{d^2 f_n(x)}{dx^2} \right]_0^\ell,$$

and using Eq. (24), the following result is obtained:

$$\frac{(\omega_m^2 - \omega_n^2)}{c^2}\Lambda_{nm} = i(\omega_m - \omega_n)\left[\frac{f_n(0)f_m(0)}{\zeta} + \frac{f_n(\ell)f_m(\ell)}{\zeta_\ell} \right].$$

For $\omega_m \neq \omega_n$, because $\omega_m+\omega_n \neq 0$, the expression of Λ_{nm} is deduced. For $\omega_m=\omega_n$, the calculation is straightforward. The general formula is found to be

$$\Lambda_{nm} = \frac{ci}{\omega_m + \omega_n}\left[\frac{f_n(0)f_m(0)}{\zeta} + \frac{f_n(\ell)f_m(\ell)}{\zeta_\ell} \right] + \frac{1}{2}\ell\,\delta_{nm}, \tag{39}$$

where δ_{nm} is the Kronecker symbol, or:

$$\Lambda_{nm} = -\frac{c}{2}\frac{\sinh 2\eta + (-1)^{n+m}\sinh 2\eta_\ell}{i(\omega_m + \omega_n)} + \frac{1}{2}\ell\,\delta_{nm}. \tag{40}$$

Modes are found to be nonorthogonal for the product defined by Eq. (38), but, as shown by Guyader,[2] it is possible to solve the problem from the knowledge of initial conditions. When dissipation tends to zero (η_r and $\eta_{\ell r}$ tend to zero), the first term does not vanish, tending to $(-1)^\mu \frac{1}{2}\ell\,\delta_{n,\nu}$. This is due to the choice of considering separately the modes ω_n and ω_ν.

Otherwise, formula (39) remains valid when one of the modes is the constant mode $f(x)=1$, and the other one a non constant mode:

$$\Lambda_n = \int_0^\ell f_n(x)dx = -\frac{c}{i\omega_n}\left[\frac{f_n(0)}{\zeta} + \frac{f_n(\ell)}{\zeta_\ell} \right] = -\frac{c}{i\omega_n}(\sinh \eta + (-1)^n \sinh \eta_\ell). \tag{41}$$

Finally, the product of the constant mode by itself is ℓ.

3. Solution with respect to initial conditions

According to Eq. (37), the initial conditions are:

$$p(x,0) = \sum_n A_n \cosh(i\omega_n x/c + \eta) + A; \tag{42}$$

$$\partial_t p(x,0) = \sum_n A_n i\omega_n \cosh(i\omega_n x/c + \eta). \tag{43}$$

Using Eq. (40) for a nonconstant mode m, the following results are obtained:

$$\int_0^\ell p(x,0) f_m(x) dx = \sum_n A_n \Lambda_{nm} + A\Lambda_m; \tag{44}$$

$$\int_0^\ell \partial_t p(x,0) f_m(x) dx = \sum_n A_n i\omega_n \Lambda_{nm}. \tag{45}$$

Multiplying Eq. (44) by $i\omega_m$, then adding Eq. (45), leads to

$$\int_0^\ell [i\omega_m p(x,0) + \partial_t p(x,0)] f_m(x) dx = i\sum_n A_n(\omega_m + \omega_n)\Lambda_{nm} + iA\omega_m\Lambda_m = -c\hat{\sum_n} A_n\left[\frac{f_n(0)f_m(0)}{\zeta} + \frac{f_n(\ell)f_m(\ell)}{\zeta_\ell} \right] + iA_m\ell\omega_m, \tag{46}$$

$$= -c\left[\frac{f_m(0)p(0,0)}{\zeta} + \frac{f_m(\ell)p(\ell,0)}{\zeta_\ell}\right] + iA_m\ell\omega_m. \tag{47}$$

Notation $\hat{\Sigma}$ for the series in Eq. (46) indicates that it involves the constant mode. As noticed by Guyader,[2] this series is related to the initial conditions at the two ends $x=0$ and $x=\ell$. Thus for a nonconstant mode:

$$A_n\ell i\omega_n = \int_0^\ell [i\omega_n p(x,0) + \partial_t p(x,0)] f_n(x)dx + cp(0,0)\sinh\eta + cp(\ell,0)\varepsilon_n \sinh\eta_\ell. \tag{48}$$

The following property is deduced from Eq. (27): $A_\nu f_\nu(x) = A_n^* f_n^*(x)$, thus $p(x,t)$ is real. Calculating $\int_0^\ell \partial_t p(x,0)dx$, we similarly get coefficient A:

$$A = \frac{c^{-1}\int_0^\ell \partial_t p(x,0)dx + p(0,0)\tanh\eta + p(\ell,0)\tanh\eta_\ell}{\tanh\eta + \tanh\eta_\ell}. \tag{49}$$

What is the condition for which this coefficient vanishes? If for instance at $x=0$, ζ is zero, η is infinite, and, according to the boundary condition, $p(0,0)$ vanishes, thus A vanishes too. This confirms the remark concerning result (26).

Using the initial conditions for the Green function found in Eq. (17), it is possible to check result (26), but this will be done hereafter using the equation with source.

B. First-order system of equations, orthogonality, and completeness of the modes

1. Introduction

In this section, we will prove that the modes form a Riesz basis in the space of solutions of a closely related problem, and give the expression of a scalar product making the modes orthogonal. As an introduction, we show that a modified scalar product leads to the orthogonality of modes, except the constant one. For vibrating systems, the product defined by Eq. (38) corresponds to the product with respect to the mass, a complement being the calculation of the product related to the stiffness (see, e.g., Meirovitch):[20]

$$\Lambda'_{nm} = \int_0^\ell \frac{d}{dx}f_n(x)\frac{d}{dx}f_m(x)dx.$$

By integrating by parts, and using Eq. (39), this product, for $n \neq m$, is found to be equal to:

$$\Lambda'_{nm} = [f_n(x)d_x f_m(x)]_0^\ell + \frac{\omega_m^2}{c^2}\Lambda_{nm} = -\frac{\omega_n\omega_m}{c^2}\Lambda_{nm}.$$

Therefore, the modes become orthogonal if we define a new product, as follows:

$$\int_0^\ell \left[\partial_x p_n \partial_x p_m - \frac{1}{c^2}\partial_t p_n \partial_t p_m\right]_{t=0} dx = \delta_{nm}\ell\omega_n^2/c^2, \tag{50}$$

where $p_n = p_n(x,t) = f_n(x)\exp(i\omega_n t)$ and similarly for index m. We remark that the modes p_n and $p_\nu = (-1)^\mu p_n^*$ are orthogonal for this product. For the calculation of the solution from initial conditions, using Eq. (37) at $t=0$, the following result is obtained:

$$\int_0^\ell \left[\frac{d}{dx}f_n(x)\partial_x p(x,0) - \frac{i\omega_n}{c^2}f_n(x)\partial_t p(x,0)\right]dx = A_n\frac{\omega_n^2}{c^2}\ell. \tag{51}$$

As a consequence, the initial conditions need to be written by using the derivatives of the function $p(x,t)$ with respect to abscissa and time, respectively. Result (48) can be checked by integrating by parts the first term of the integral. Nevertheless, the product (50) is not useful for the constant mode, and the first method needs to be used (see Subsection VI A 3). Moreover, this derivation does not prove that the product is a scalar product, and that the modes form a basis of the space of solutions of the problem. This will be done hereafter.

2. Riesz basis of the modes

Several works have been done by mathematicians concerning spectral operators when boundary conditions are not simple conditions, such as Neumann or Dirichlet conditions. We quote the work by Russell,[21] Majda,[22] Lagnese,[23] Banks *et al.*,[24] Darmawijoyo and Van Horssen,[7] and Cox and Zuazua.[6] Rideau[4] has treated the 1D case with a (unique) resistive termination, giving explicitly a scalar product (see also Ref. 5). We generalize his calculation using a similar method, by considering the wave equation with source in the following form:

$$\partial_t \boldsymbol{\psi}(x,t) = \mathbf{A}\boldsymbol{\psi}(x,t) + \boldsymbol{\phi}_s(x,t), \tag{52}$$

where $\boldsymbol{\psi}(x,t) = (p,v)^T$, p and $v/(\rho c)$ being the acoustic pressure and velocity, respectively. Operator $\mathbf{A}$ is:

$$\mathbf{A} = \begin{pmatrix} 0 & -c\partial_x \\ -c\partial_x & 0 \end{pmatrix}, \tag{53}$$

and boundary conditions are written as:

$$p(0,t) = -\zeta v(0,t) \text{ and } p(\ell,t) = \zeta_\ell v(\ell,t)\ \forall\, t. \tag{54}$$

The family of eigenelements of $\mathbf{A}$ are found to satisfy:

$$\lambda_n p_n(x) = -c\partial_x v_n(x); \quad \lambda_n v_n(x) = -c\partial_x p_n(x), \tag{55}$$

thus

$$\begin{pmatrix} p_n(x) \\ v_n(x) \end{pmatrix} = \begin{pmatrix} \cosh(\lambda_n x/c + \eta) \\ -\sinh(\lambda_n x/c + \eta) \end{pmatrix} \tag{56}$$

$$\lambda_n = (-\eta - \eta_\ell + in\pi)c/\ell = i\omega_n \tag{57}$$

[see Eq. (20)]. $p_n(x) = f_n(x)$ and λ_n are identical to the eigenfunctions and eigenvalues found before. Nevertheless, the constant mode is eliminated (except for the very particular case $\eta = -\eta_\ell$), because the boundary conditions are slightly different: Eqs. (2) and (3) are obtained by deriving Eq. (54) with respect to t. In Eq. (56), the argument of the hyperbolic functions can be written as:

$$\lambda_n x/c + \eta = \alpha(x) + i\beta_n(x); \tag{58}$$

$$\alpha(x) = -\eta_{\ell r} x/\ell + \eta_r(1 - x/\ell); \tag{59}$$

$$\beta_n(x) = \pi[-\mu_\ell x/\ell + \mu(1 - x/\ell)]/2 + n\pi x/\ell. \tag{60}$$

[See definitions (10) and (11)]. Denoting $\boldsymbol{\psi}_n^\alpha(x) = (p_n(x), v_n(x))^T$, we show in the Appendix that the family of elements $\boldsymbol{\psi}_n^\alpha(x)$ is a Riesz basis, i.e., a complete basis of elements, which become orthogonal for the following scalar product:

$$\langle \boldsymbol{\psi}, \boldsymbol{\varphi} \rangle_H^\alpha = \int_0^\ell \boldsymbol{\varphi}^{T^*} \mathbf{G}_{2\alpha}(x) \boldsymbol{\psi} dx, \tag{61}$$

$$\text{where } \mathbf{G}_\alpha(x) = \begin{pmatrix} \cosh \alpha(x) & \sinh \alpha(x) \\ \sinh \alpha(x) & \cosh \alpha(x) \end{pmatrix}. \tag{62}$$

For a given vector $\boldsymbol{\psi} = (p, v)^T$, the value of the scalar product with eigenvector $\boldsymbol{\psi}_n^\alpha$ is found to be:

$$\langle \boldsymbol{\psi}, \boldsymbol{\psi}_n^\alpha \rangle_H^\alpha = \int_0^\ell [p(x,t)p_n(x) - v(x,t)v_n(x)] dx. \tag{63}$$

This is in accordance with the product (50). A direct application of this result is the solution of Eq. (52) with initial conditions $\boldsymbol{\psi}(x,0) = [p(x,0), v(x,0)]^T$. The modal decomposition is uniquely determined as $\boldsymbol{\psi}(x,t) = \Sigma_n h_n(t) \boldsymbol{\psi}_n^\alpha(x)$ in the energy space H, and leads to the following family of *decoupled* ordinary differential equations:

$$\ell[\partial_t h_n - \lambda_n h_n] = \langle \boldsymbol{\phi}_s(x,t), \boldsymbol{\psi}_n^\alpha \rangle_H^\alpha; \tag{64}$$

$$h_n(0) = \langle \boldsymbol{\psi}(x,0), \boldsymbol{\psi}_n^\alpha \rangle_H^\alpha. \tag{65}$$

This result can be first applied to the calculation done in Sec. VI A. In order to find a solution $\chi(x,t)$ of the second-order equation without source, we denote $\boldsymbol{\psi}(x,t) = [\partial_t \chi(x,t), -c\partial_x \chi(x,t)]^T$, and obtain by integrating $\boldsymbol{\psi}(x,t)$ with respect to t:

$$\chi(x,0) = \sum_n \lambda_n^{-1} h_n(0) p_n(x) + A. \tag{66}$$

Using Eqs. (65) and (63), and replacing $\chi(x,t)$ by $p(x,t)$, $p_n(x)$ by $f_n(x)$, λ_n by $i\omega_n$, and $\lambda_n^{-1} h_n(0)$ by A_n, formula (51) is checked. Notice that coefficient A cannot be directly determined with this method.

3. Example of the Green function

Similarly, the Green function can be calculated by using the previous result. In order for the unknown function to satisfy the boundary conditions (54), or (2) and (3), it is convenient to define the following vectors:

$$\boldsymbol{\psi}(x,t) = \begin{pmatrix} \partial_t g(x,t) \\ -c\partial_x g(x,t) \end{pmatrix}; \quad \boldsymbol{\phi}_s(x,t) = \begin{pmatrix} c^2 \delta(x - x_0)\delta(t) \\ 0 \end{pmatrix}. \tag{67}$$

The first row of Eq. (52) is Eq. (1), while the second one comes from the definition of vector $\boldsymbol{\psi}$. Using Eq. (64), the solution is found to be: $\boldsymbol{\psi}(x,t) = \Sigma_n \boldsymbol{\psi}_n^\alpha(x) h_n(t)$, where

$$\partial_t h_n - \lambda_n h_n = c^2 \ell^{-1} p_n(x_0) \delta(t). \tag{68}$$

The initial conditions for the Green function imply $\boldsymbol{\psi}(x,t) = 0$ for $t < 0$, therefore $h_n(t) = 0$ for $t < 0$. Thus, the solution of Eq. (68) is

$$h_n(t) = A_n e^{\lambda_n t} H(t); \quad A_n = c^2 \ell^{-1} p_n(x_0).$$

As a consequence,

$$\begin{pmatrix} \partial_t g(x,t) \\ -c\partial_x g(x,t) \end{pmatrix} = \sum_n A_n \begin{pmatrix} p_n(x) \\ v_n(x) \end{pmatrix} e^{\lambda_n t} H(t). \tag{69}$$

Integrating the first row with respect to time leads to:

$$g(x,t) = H(t) \sum_n [A_n p_n(x) e^{\lambda_n t} + A(x)]. \tag{70}$$

Derivating this expression with respect to x and using the second row of Eq. (69) leads to $\partial_x A(x) = 0$, thus A is a constant, as expected. In order to deduce the value of this constant, we need the following result:

$$\partial_t g(x,0) = p(x,0) = c^2 \delta(x - x_0). \tag{71}$$

It is obtained by derivating the first row of Eq. (69) with respect to time, and the second row of Eq. (69) with respect to abscissa, leading to $p(x,0)\delta(t) = c^2 \delta(x - x_0)\delta(t)$ [recall that $\partial_t[F(t)H(t)] = H(t)\partial_t F(t) + F(0)\delta(t)$]. The end of the calculation is done in Sec. VI A 3, giving Eq. (49), by replacing $p(x,0)$ by $g(x,0)$ and taking into account that $g(0,0) = g(\ell,0) = 0$. We notice that the calculation is valid for both passive and active boundaries.

VII. EIGENMODES EXPANSION IN THE FREQUENCY DOMAIN: BIORTHOGONALITY

Frequency domain approach is very popular in acoustics (see, e.g., Ref. 7), and leads to the use of biorthogonality (see, e.g., Ref. 16, p. 884) of modes, except when the boundary impedances are imaginary, corresponding to nondissipative boundaries: For that case, modes are orthogonal, and the Laplacian operator is self-adjoint. In this section, we limit the discussion to the Green function calculation, and successively use the two above-used approaches: The second-order equation, then the system of two first-order equations, ignoring the constant mode. Because we are now in the Fourier domain, equations are ordinary differential equations, biorthogonality theory ensuring the completeness of the modes family.

A. Solution of the second-order equation

1. Modal expansion

In order to calculate the inverse FT of $G(x,\omega)$, another solution is possible: The expansion of $G(x,\omega)$ in eigenmodes. This is done for a particular case by Filippi (Ref. 1 p. 58: This author considers another type of excitation instead of the Dirac function, thus uses the Laplace transform instead of the FT; notice that the constant mode is missing in this

work). We will see how this method leads to the same poles and residues that the direct method using the closed-form expression (12). We are searching for the following expansion:

$$G(x,\omega)=\sum_n G_n(x,\omega), \tag{72}$$

where the eigenmodes $G_n(x,\omega)$ are solutions of:

$$[\partial_{xx}^2+\theta_n^2(\omega)/c^2]G_n(x,\omega)=0, \tag{73}$$

and satisfy the boundary conditions (7) and (8). The key point is that eigenmodes $G_n(x,\omega)$ and eigenfrequencies $\theta_n(\omega)$ depend on frequency ω: This is due to the boundary conditions, which are of Robin type. Solutions of Eq. (73) can be written as follows:

$$G_n(x,\omega)=\cosh(i\theta_n(\omega)x/c+\varphi_n(\omega)), \tag{74}$$

where $\theta_n(\omega)$ and $\varphi_n(\omega)$ are given by the boundary conditions. Thus, they satisfy

$$\theta_n(\omega)\tanh\varphi_n(\omega)=\omega/\zeta; \tag{75}$$

$$\theta_n(\omega)\tanh(i\theta_n(\omega)\ell/c+\varphi_n(\omega))=-\omega/\zeta_\ell. \tag{76}$$

Eliminating quantity $\varphi_n(\omega)$, the eigenvalues are found to satisfy the following equation:

$$\tanh(i\theta_n(\omega)\ell/c)\left[\theta_n(\omega)+\frac{\omega^2}{\theta_n(\omega)\zeta\zeta_\ell}\right]=-\omega\left[\frac{1}{\zeta}+\frac{1}{\zeta_\ell}\right]. \tag{77}$$

When $\theta_n(\omega)$ and ω are not simultaneously zero, this equation can be rewritten as

$$e^{2i\theta_n\ell/c}=\left[\frac{\theta_n\zeta-\omega}{\theta_n\zeta+\omega}\right]\left[\frac{\theta_n\zeta_\ell-\omega}{\theta_n\zeta_\ell+\omega}\right]. \tag{78}$$

Calculation of all solutions of this equation is not necessary, only one of them being useful in the following. Operator $D=\partial_{xx}^2$ is formally equal to its adjoint $\bar{D}$, but the boundary conditions are different (conditions for $\bar{D}$ are complex conjugate of conditions for D). Modes of $\bar{D}$ are the complex conjugate of modes $G_n(x,\omega)$ [they are equal to modes $G_n(x,\omega)$ only if ζ and ζ_ℓ are imaginary, because of the factor i in boundary conditions (7) and (8)]. Thus, in general, operator D is not self-adjoint, and eigenmodes $G_n(x,\omega)$ and $\overline{G_n}(x,\omega)=G_n^*(x,\omega)$ are biorthogonal (see Ref. 16). The scalar product of modes $G_n(x,\omega)$ with $\overline{G_m}(x,\omega)$ is simply given by:

$$\int_0^\ell G_m(x,\omega)G_n(x,\omega)dx=\Gamma_n\delta_{nm}, \tag{79}$$

where

$$\Gamma_n=\frac{\ell}{2}\left[1+c\frac{\sinh 2(i\theta_n\ell/c+\varphi_n)-\sinh 2\varphi_n}{2i\ell\theta_n}\right]. \tag{80}$$

Therefore modes $G_n(x,\omega)$ are orthogonal [for the product (79)] and fulfill the same boundary conditions as $G(x,\omega)$, contrary to resonance modes $f_n(x)$ in Eq. (29). Finally, the solution of Eq. (6) can be written as follows:

$$G(x,\omega)=c^2\sum_n\frac{G_n(x,\omega)G_n(x_0,\omega)}{\Gamma_n(\theta_n^2(\omega)-\omega^2)}. \tag{81}$$

2. Calculation of poles and residues

In order to calculate the inverse FT, the residue calculus will be used again. The only terms of the series contributing to poles verify:

$$\theta_n(\omega)=\pm\omega. \tag{82}$$

Looking at Eq. (78), it can be seen that these two solutions lead to the same equation for ω. Rewriting Eq. (78) by using Eqs. (75) and (76), the resonance modes frequencies are found to be solutions of Eq. (20). Solutions ω_p of this equation are the nonzero poles of the integral in the inverse FT. Nevertheless, the pole $\omega=0$ exists again, because the zero value satisfies Eq. (82), the eigenvalue $\theta_n(\omega)=0$ satisfying Eq. (77).

It remains to calculate the residues. Starting with the poles $\omega_p\neq 0$, we need to select in the series (81) the terms involving poles. For a given ω_p, there are two terms. However, it appears that modes corresponding to θ_n and $-\theta_n$ are identical. As a consequence, only one term of the series contributes to the inverse FT: It will be denoted $\theta_p(\omega)$. The corresponding residue is found by expanding Eq. (81) for ω close to ω_p, as follows:

$$G(x,\omega)=c^2\frac{G_p(x,\omega_p)G_p(x_0,\omega_p)}{\Gamma_p(2\omega_p)(\omega_p-\omega)\left(1-\left[\frac{d}{d\omega}\theta_p(\omega)\right]_{\omega=\omega_p}\right)}.$$

A similar expression can be found in Filippi,[1] which points out that Morse and Ingard (Ref. 7 p. 559) forgot the derivative. The same error is found in Morse and Feshbach (Ref. 16 p. 1347), with another error in the derivation of Eq. (77): These authors treated the problem of a string with one nonrigid (and resistive) support.

Actually, the derivative of $\theta_p(\omega)$, denoted $\theta_p'(\omega)$ can be calculated analytically, as follows. Taking the derivative of Eq. (78) with respect to ω, or more conveniently, taking the logarithmic derivative of Eqs. (75) and (76), the following results are obtained:

$$\frac{1}{\omega_p}-\frac{\theta_p'}{\theta_p}=\frac{2\varphi_p'}{\sinh 2\varphi_p}=\frac{2(i\theta_p'\ell/c+\varphi_p')}{\sinh 2(i\theta_p\ell/c+\varphi_p)}.$$

Thus, eliminating the derivative φ_p', writing $\theta_p=\omega_p$ and using Eq. (80), it is found after some algebra:

$$1-\theta_p'=\ell/2\Gamma_p. \tag{83}$$

Finally, for ω close to ω_p:

$$G(x,\omega)=-\frac{c^2}{\ell}\frac{G_p(x,\omega_p)G_p(x_0,\omega_p)}{\omega_p(\omega-\omega_p)}, \tag{84}$$

which is in accordance with Eq. (23).

Otherwise, for ω close to 0, the solution $\theta(\omega)$ which is close to 0, solution of Eq. (82), satisfies the following equation, deduced from Eq. (77):

$$[1-\Theta^2/3+O(\Theta^4)][\Theta^2+\Omega^2/(\zeta\zeta_\ell)]=i\Omega[\zeta^{-1}+\zeta_\ell^{-1}],$$

where $\Theta=\theta\ell/c$ and $\Omega=\omega\ell/c$. Therefore, Θ^2 is of order Ω, and

$$\theta^2=i\omega c\ell^{-1}[\zeta^{-1}+\zeta_\ell^{-1}]+O(\omega^2). \tag{85}$$

Using Eq. (81), the residue for the pole $\omega=0$ is obtained, and Eq. (25) is confirmed. We conclude that the method of the expansion of orthogonal modes in the Fourier domain leads to the same result [Eq. (26)] than the "direct" method, but the derivation is more delicate.

B. System of two first-order equations

We consider now the FT of Eq. (52):

$$i\omega\boldsymbol{\psi}(x,\omega)=\mathbf{A}\boldsymbol{\psi}(x,\omega)+\boldsymbol{\phi}_s(x,\omega), \tag{86}$$

where $\boldsymbol{\psi}(x,\omega)=[P(x,\omega),V(x,\omega)]^T$. It is interesting that the boundary conditions are independent of frequency:

$$P(0,\omega)=-\zeta V(0,\omega);\quad P(\ell,\omega)=\zeta_\ell V(\ell,\omega). \tag{87}$$

Eigenvalues λ_n and eigenvectors $\boldsymbol{\psi}_n^\alpha$ of operator $\mathbf{A}$ are already known [see Eqs. (56) and (57)]. The Appendix shows that the adjoint operator of $\mathbf{A}$, is $\bar{\mathbf{A}}=-\mathbf{A}$, and gives the boundary conditions for it. This formulation differs slightly from the work,[9,10] considering a different operator, but the principle is identical: We notice that these authors treat the problem for more general operators and boundary conditions. Eigenvalues and eigenvectors of $\bar{\mathbf{A}}$ are solutions of:

$$\bar{\mathbf{A}}\,\overline{\boldsymbol{\psi}_m^\alpha}=\overline{\lambda_m}\,\overline{\boldsymbol{\psi}_m^\alpha},$$

$$\overline{p_m}=\zeta\overline{v_m}\ \text{for}\ x=0\ \overline{p_m}=-\zeta_\ell\overline{v_m}\ \text{for}\ x=\ell.$$

Thus, the adjoint eigenvalue problem to be solved is the same as the direct one, by replacing c by $-c$, η by $-\eta$ and η_ℓ by $-\eta_\ell$. The eigenelements are thus found to be:

$$\overline{\boldsymbol{\psi}_m^\alpha}=\begin{pmatrix}\overline{p_m}(x)\\ \overline{v_m}(x)\end{pmatrix}=\begin{pmatrix}\cosh(\overline{\lambda_m}x/c+\eta)\\ \sinh(\overline{\lambda_m}x/c+\eta)\end{pmatrix}, \tag{88}$$

where

$$\overline{\lambda_m}=-(\eta+\eta_\ell+im\pi)c/\ell=\lambda_{-m}. \tag{89}$$

Comparing with the family $\boldsymbol{\psi}_n^\alpha$ [Eq. (56)], there is a difference in sign for the second row: We notice that Rideau[3] made an error in the biorthogonal family. By construction, the biorthogonality relationship is ensured:

$$(\lambda_n-\overline{\lambda_m}^*)<\boldsymbol{\psi}_n^\alpha,\quad\overline{\boldsymbol{\psi}_m^\alpha}>=0. \tag{90}$$

Using Eq. (27), we remark that $\lambda_n=\overline{\lambda_m}^*$ implies $m=-\nu$, as defined in Eq. (28). Therefore

$$\langle\boldsymbol{\psi}_n^\alpha,\overline{\boldsymbol{\psi}_m^\alpha}\rangle=\int_0^\ell\overline{\boldsymbol{\psi}_m^{\alpha T*}}\boldsymbol{\psi}_n^\alpha dx=(-1)^\mu\ell\,\delta_{m,-\nu}. \tag{91}$$

This latter relation enables one to perform a modal decomposition on the $(\boldsymbol{\psi}_n^\alpha)$ family: But, contrarily to standard cases, the nth coefficient is not given by the scalar product with $\boldsymbol{\psi}_n^\alpha$, but by the scalar product with $\boldsymbol{\psi}_{-\nu_n}^\alpha$ [up to the normalization coefficient $(-1)^\mu\ell$]. Notice that if η and η_ℓ are both real, $\overline{\lambda_m}=\lambda_m^*$, and Eq. (91) is obvious from the expressions of eigenvalues and eigenvectors (for this case, $\lambda_n=\overline{\lambda_m}^*$ implies $n=m$). For the general case, the scalar product can be written: $\langle\boldsymbol{\psi}_n^\alpha,\overline{\boldsymbol{\psi}_m^\alpha}\rangle=\int_0^\ell\cos[\beta_n(x)+\beta_{-m}(x)]dx=(-1)^\mu\int_0^\ell\cos[(\nu+m)\pi x/\ell]dx$. Comparison with Eq. (A4) exhibits the difference between the two methods.

It remains to apply orthogonality to Eq. (86). We choose the case of the Green function [Eq. (67)], with the following result:

$$G(x,\omega)=-\frac{c^2}{\ell}\sum_n\frac{f_n(x)f_x(x_0)}{\omega(\omega-\omega_n)}. \tag{92}$$

The calculation is easy, because ω_n does not depend on frequency. Comparison with Eq. (29) exhibits a difference in the denominator, i.e., a factor ω instead of ω_n, and, of course, the absence of constant mode. When returning to the time domain, all of the terms corresponding to $\omega=\omega_n$ are identical, and a constant mode is found for the pole $\omega=0$, but again it is not possible to deduce it from orthogonality relations, as in Sec. VI B. Nevertheless, because of the independence of the boundary conditions with respect to frequency, the calculation of the residues is much easier than for the second-order equation. For the same reason, the calculation in the time domain would be possible with the same modal decomposition, and this is a major difference with the methods based upon the second-order equation.

VIII. CONCLUSION

The simple problem we have studied, which can be regarded in particular as a radiation problem, exhibits interesting properties for the resonance modes: They are complex valued, and nonorthogonal for the simple product (38) because of the bounded character of the considered medium, but except the constant mode, they are orthogonal for a product modified in a proper way, and are a basis for the space of solutions. Second-order equations allow one to find the constant mode, while first-order systems of equations allow a more direct formulation of boundary impedances.

Thanks to the simplicity of the problem, the analytical treatment is possible with several methods, elucidating the relationship between them, which can be useful for more intricate problems (e.g., when damping is added to propagation, or when boundary impedances involve a mass or a spring). No approximations are needed, the results are valid whatever the value of the terminal resistances. Active boundaries can also be considered, thanks to a change in functions. We notice that an advantage of the frequency domain calculations is the possibility of the treatment of an arbitrary dependence of the boundary conditions. For a dependence $\eta(\omega)$ and $\eta_\ell(\omega)$, Eq. (29) remains valid by replacing ℓ by $[\ell$

$-ic(\eta'_p+\eta'_{\ell p})]$, where $\eta'_p=(d\eta/d\omega)_{\omega=\omega_p}$, and similarly for η_ℓ. This can be shown by generalizing Eq. (22), or, with some algebra, using the modal expansion.

Finally, considering the problem of a stratified medium (see Sec. II), it could be deduced in the field outside of the interval $[0,\ell]$. When terminations are passive, a result is, that modes tend to infinity when x tends to $\pm\infty$. An interesting study has been done in Ref. 13, using biorthogonality and explaining the relation between the energy outside of the interval and the terms responsible of nonorthogonality in Eq. (39).

ACKNOWLEDGMENTS

We would like to thank José Antunes, Patrick Ballard, Sergio Bellizzi, Michel Bruneau, Paul Filippi, Dominique Habault, Pierre-Olivier Mattei, and Vincent Pagneux for very fruitful discussions.

APPENDIX: PROOF OF THE COMPLETENESS OF THE EIGENELEMENTS OF OPERATOR A

The operator $\mathbf{A}$ defined by Eq. (53) is a differential operator defined on the energy space $H=L^2(0,\ell)\times L^2(0,\ell)$, it has a compact resolvent (cf Ref. 4, p. 191). Using the ordinary scalar product $\langle\boldsymbol{\psi},\boldsymbol{\varphi}\rangle=\int_0^\ell[pq^*+vw^*]dx$, between $\boldsymbol{\psi}(x,t)=(p,v)^T$ and $\boldsymbol{\varphi}(x,t)=(q,w)^T$, the following result is obtained:

$$\langle\mathbf{A}\boldsymbol{\psi},\boldsymbol{\varphi}\rangle+\langle\boldsymbol{\psi},\mathbf{A}\boldsymbol{\varphi}\rangle=-c[v(q^*+\zeta_\ell w^*)]_{x=\ell}+c[v(q^*-\zeta w^*)]_{x=0}.$$

It is deduced that the adjoint operator of $\mathbf{A}$ is $\bar{\mathbf{A}}=-\mathbf{A}$ (we denote all quantities related to the adjoint problem with an overline), and on its domain, the following adjoint boundary conditions must be fulfilled:

$$q(0,t)=\zeta^* w(0,t) \text{ and } q(\ell,t)=-\zeta_\ell^* w(\ell,t)\ \ \forall\, t.$$

[Here, $\zeta=\zeta^*$, and $\zeta_\ell=\zeta_\ell^*$; if ζ is infinite, the boundary conditions are $v(0,t)=0$, and $w(0,t)=0$, and similarly for boundary $x=\ell$]. Therefore, $\mathbf{A}$ is skew *symmetric*, but not skew *adjoint*, because the domains of $\mathbf{A}$ and $\bar{\mathbf{A}}$ are different, except if both ζ or ζ_ℓ are either zero or infinite (Dirichlet or Neumann conditions); notice that for a skew-adjoint operator, the eigenvalues are purely imaginary. In order to find a new scalar product, we denote, from Eqs. (58)–(60):

$$\boldsymbol{\psi}_n^\alpha(x)=\begin{pmatrix} e^{\alpha(x)} & e^{-\alpha(x)} \\ -e^{\alpha(x)} & e^{-\alpha(x)}\end{pmatrix}\begin{pmatrix} e^{i\beta_n(x)} \\ e^{-i\beta_n(x)}\end{pmatrix}.$$

In H, the standard scalar product $\langle\boldsymbol{\psi}_n^\alpha,\boldsymbol{\psi}_p^\alpha\rangle_H=\int_0^\ell[p_n p_p^*+v_n v_p^*]dx$ does not vanish for $n\neq p$, except if $\alpha(x)=0$. If we denote $\boldsymbol{\psi}_n^0(x)$, the functions corresponding to the latter case, it is possible to construct a new scalar product ensuring orthogonality, in a similar way to Rideau.[3] From Eq. (62), the following *hyperbolic rotation* is obtained:

$$\boldsymbol{\psi}_n^0(x)=\mathbf{G}_\alpha(x)\boldsymbol{\psi}_n^\alpha(x). \tag{A1}$$

We will now prove that the new product

$$\langle\boldsymbol{\psi},\boldsymbol{\varphi}\rangle_H^\alpha=\langle\mathbf{G}_\alpha\boldsymbol{\psi},\mathbf{G}_\alpha\boldsymbol{\varphi}\rangle_H=\int_0^\ell\boldsymbol{\varphi}^{T*}\mathbf{M}_\alpha(x)\boldsymbol{\psi}dx, \tag{A2}$$

where $\mathbf{M}_\alpha(x)=\mathbf{G}_\alpha^T\mathbf{G}_\alpha$, leads to the orthogonality of the modes. $\mathbf{M}_\alpha(x)$ is found to be equal to $\mathbf{G}_{2\alpha}$. It is symmetrical and positive definite because

$$(\|\boldsymbol{\psi}\|_H^\alpha)^2=(\|(p,v)^T\|_H^\alpha)^2=\int_0^\ell[\cosh[2\alpha(x)](|p|^2+|v|^2)+2\sinh[2\alpha(x)]\Re e(pv^*)]dx$$

can be rewritten as:

$$(\|\boldsymbol{\psi}\|_H^\alpha)^2=\frac{1}{2}\int_0^\ell[e^{2\alpha(x)}|p+v|^2+e^{-2\alpha(x)}|p-v|^2]dx.$$

Moreover, $\alpha(x)$ is a function varying monotonously from η_r to $-\eta_{\ell r}$ when x increases from 0 to ℓ, and the following bounds can be found for $\|\boldsymbol{\psi}\|_H^\alpha$:

$$c_\alpha\|\boldsymbol{\psi}\|_H<\|\boldsymbol{\psi}\|_H^\alpha<C_\alpha<\|\boldsymbol{\psi}\|_H, \tag{A3}$$

where $c_\alpha=e^{-\tilde{\eta}}$ and $C_\alpha=e^{\tilde{\eta}}$, with $\tilde{\eta}=\sup[|\eta_r|,.|\eta_{\ell r}|]$.

Therefore the modes $\boldsymbol{\psi}_n^\alpha$ are orthogonal for the new scalar product $\langle\boldsymbol{\psi},\boldsymbol{\varphi}\rangle_H^\alpha$. First, recall that $(\boldsymbol{\psi}_n^0)_n$ is the family of eigenvectors of a classically skew-adjoint operator with compact resolvent, it is thus complete in H. Now, thanks to Eq. (A3), the two norms are equivalent on H, and the hyperbolic rotation shows that $(\boldsymbol{\psi}_n^0)_n$ and $(\boldsymbol{\psi}_n^\alpha)_n$ span the same subspace, namely the whole of H. This proves the completeness of $(\boldsymbol{\psi}_n^\alpha)_n$ in H.

The calculation leads to the simple result:

$$\langle\boldsymbol{\psi}_n^\alpha,\boldsymbol{\psi}_m^\alpha\rangle_H^\alpha=\langle\boldsymbol{\psi}_n^0,\boldsymbol{\psi}_m^0\rangle_H=\int_0^\ell\cos[\beta_n(x)-\beta_m(x)]dx=\int_0^\ell\cos[(n-m)\pi x/\ell]dx=\ell\,\delta_{nm}. \tag{A4}$$

[1] P. Filippi, "Acoustics of enclosures," in *Acoustics: Basic Physics, Theory and Methods*, edited by P. Filippi, D. Habault, J. P. Lefebvre, and A. Bergassoli (Academic, San Diego, 1999), Chap. 2.

[2] G. Oliveto and A. Santini, "Complex modal analysis of a continuous model with radiation damping," J. Sound Vib. **192**, 15–33 (1996).

[3] J. L. Guyader, *Vibrations des Milieux Continus (Vibration of Continuous Media)* (Lavoisier, Hermès, Paris, 2002).

[4] P. Rideau, "Contrôle par assemblage de poutres flexibles par des capteurs-actionneurs ponctuels: Étude du spectre du système," (Control by assembling of flexible beams using discrete sensors-actuators: Study of the spectrum of the system), Ph.D. thesis, Ecole Nationale Supérieure des Mines de Paris, 1985.

[5] A. Intissar, *Analyse Fonctionnelle et Théorie Spectrale pour les Opérateurs Compacts Nonautoadjoints (Functional Analysis and Spectral Theory for Compact, Nonself-adjoint Operators)* (Cepadues, Toulouse, 1997).

[6] S. Cox and E. Zuazua, "The rate at which energy decays in a string damped at one end," Indiana Univ. Math. J. **44**, 545–573 (1995).

[7] Darmawijoyo and W. T. Van Horssen, "On the weakly damped vibrations of a string attached to a spring mass dashpot system," J. Vib. Control **9**, 1231–1248 (2003).

[8] P. M. Morse and K. U. Ingard, *Theoretical Acoustics* (McGraw–Hill, New York, 1968).

[9]M. Ochmann and U. Donner, "Investigations of silencers with asymmetrical lining. I: Theory," Acta Acust. **2**, 247–255 (1994).
[10]R. Rabenstein and L. Trautmann, "Multidimensional transfer function models," IEEE Trans. Circuits Syst. **49**, 852–861 (2002).
[11]L. Trautmann and R. Rabenstein, *Digital Sound Synthesis by Physical Modeling Using the Functional Transformation Method* (Kluwer, New York, 2003).
[12]H. Levine, "Acoustical cavity excitation," J. Acoust. Soc. Am. **109**, 2555–2565 (2001).
[13]P. T. Leung, S. Y. Liu, and K. Young, "Completeness and orthogonality of quasinormal modes in leaky optical cavities," Phys. Rev. A **49**, 3057–3067 (1994).
[14]P. T. Leung, W. M. Suen, C. P. Sun, and K. Young, "Waves in open systems via a biorthogonal basis," Phys. Rev. E **57**, 6101–6104 (1998).
[15]J. Kergomard, "Elementary considerations on reed-instruments oscillations," in *Mechanics of Musical Instruments*, edited by A. Hirschberg, J. Kergomard, and G. Weinreich (Springer, Wien, 1995).
[16]V. Debut, "Deux études d'un instrument de musique de type clarinette: Analyse des fréquences propres du résonateur et calcul des auto-oscillations par décomposition modale" "(Two studies of a clarinetlike musical instrument: Analysis of the resonator eigenfrequencies and calculation of self-sustained oscillations using modal expansion)," Ph.D. thesis, Université Aix-Marseille II, 2004.
[17]P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw–Hill, New York, 1953).
[18]J. Woodhouse, "Idealized models of a bowed string," Acustica **79**, 233–250 (1993).
[19]K. M. Ahmida, and J. R. F.Arruda , "A comparison of time domain boundary conditions for acoustic waves in waveguides," J. Sound Vib. **255**, 663–684 (2002).
[20]L. Meirovitch, *Elements of Vibration Analysis* (McGraw–Hill, New York, 1975).
[21]D. L. Russell, "Control theory of hyperbolic equations related to certain questions in harmonic analysis and spectral theory," J. Math. Anal. Appl. **40**, 336–368 (1972).
[22]A. Majda, "The location of the spectrum for the dissipative acoustic operator," Indiana Univ. Math. J. **25**, 973–987 (1976).
[23]J. Lagnese, "Decay solutions of wave equations in a bounded region with boundary dissipation," J. Diff. Eqns. **50**, 163–182 (1983).
[24]H. T. Banks, G. Propst, and R. J. Silcox, "A comparison of time domain boundary conditions for acoustic waves in waveguides," Q. Appl. Math. **54**, 249–265 (1996).

Three-dimensional underwater sound pressure field due to sonic boom

Douglas M. Moody
The Aerospace Corporation, Los Angeles, California 90009-2957

(Received 26 August 2005; revised 9 December 2005; accepted 13 December 2005)

Over 35 years ago, a theory was developed by Sawyers [K. N. Sawyers, J. Acoust. Soc. Am. **44**, 523–524 (1968)] to predict the underwater sound pressure field due to loading of the water surface with an idealized sonic boom. In Sawyers' theory, the sonic-boom *N* wave was taken to be uniform and of infinite extent in the cross-track direction, i.e., the horizontal direction perpendicular to the direction of motion. The resulting prediction is thus two-dimensional, depending only on the in-track position and the depth into the water. The current paper presents a generalization of Sawyers' theory from two to three spatial dimensions. Three sample problems are solved, with one of the results showing that the three-dimensional method correctly reproduces Sawyers' two-dimensional results when the surface disturbance is uniform in crosstrack. Results from the final problem provide actual predictions of underwater sound generated by a cruise missile flying at an altitude of 20 feet and a speed of 2500 ft/s. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2165069]

PACS number(s): 43.28.Mw, 43.30.Nb [RR] Pages: 1368–1372

I. INTRODUCTION

Operation of supersonic aircraft or missiles over the ocean inevitably raises questions about noise transmitted into the water as a result of sonic boom impingement at the air-water interface. It is important to be able to quantitatively predict underwater noise levels resulting from such flight operations since the noise, if excessive, may pose an unacceptable disturbance to marine wildlife. Lack of credible predictive capability may, in some instances, result in the prohibition of overwater supersonic flight operations at certain times and locations.

A theory of noise transmission into water due to a sonic boom running over the surface was given by Sawyers[1] in 1968. Sawyers treats the problem of a body traveling at a constant horizontal flight speed, V, which, although supersonic in air, is subsonic relative to the sound speed in water, c. Laboratory experiments performed by Intrieri and Malcolm[2] examined the transmission of sound into a water tank resulting from sonic booms created by small high-speed projectiles fired above the water surface. For cases with $V<c$, their results were in good agreement with Sawyers' predictions. Variation of underwater SPL with various parameters in Sawyers' formulation was examined by Sparrow,[3] and the theory has recently been generalized to the case of a wavy ocean surface by Cheng and Lee[4].

The Sawyers theory is formulated in two spatial dimensions, with the overpressure field derived as a function of depth into the water, $\bar{z}$, and distance along the flight track of the aircraft, $\bar{x}$. Both the surface disturbance and the resulting underwater sound field are assumed to be uniform and of infinite extent in the third dimension (cross-track distance, $\bar{y}$). In certain situations, this assumption may be an adequate approximation to reality, such as when the flight altitude greatly exceeds the body length. In these cases, the in-track length of the boom signature at the water surface is very much smaller than the cross-track distance to which a significant boom amplitude persists. In these cases of so-called high aspect ratio, local uniformity in the cross-track coordinate is a reasonable approximation, and Sawyers' results may be expected to hold on a local basis. However, in other situations, it is necessary to abandon this approximation and fully account for the three dimensionality in the underwater sound field. Such is the case for the flight of a supersonic sea-skimming missile flying at an altitude which is on the order of its body length. In these low-aspect-ratio cases, the in-track and cross-track extents of the surface disturbance are of comparable length. This paper supplies a treatment for such cases by generalizing the Sawyers' theory from two spatial dimensions to three.

II. PROBLEM FORMULATION AND SOLUTION

The geometry of the problem is shown in Fig. 1. The elevation view (a) shows the disturbance traveling along the surface toward negative $\bar{x}$. (Overbars are used here to denote dimensional variables in the problem. Unadorned symbols will be used to represent the corresponding dimensionless variables, which will be defined shortly). The surface of the ocean is defined by the plane $\bar{z}=0$, and depth is measured positive downward. The view looking down on the ocean surface (b) shows the cross-track direction, $\bar{y}$, running positive to the left of the flight path. In this view, the disturbance at $\bar{z}=0$, at some instant in time, is shown to be confined to a "surface footprint" of finite spatial extent in $(\bar{x},\bar{y})$. In Sawyers' paper,[1] the surface disturbance and resulting underwater sound field were taken to be uniform and of infinite extent in the $\bar{y}$ direction, which yielded a two-dimensional solution for the underwater overpressure field, $\bar{p}(\bar{x},\bar{z},\bar{t})$. The goal here is to generalize the solution procedure, thus obtaining the three-dimensional underwater field, $\bar{p}(\bar{x},\bar{y},\bar{z},\bar{t})$, resulting from a given surface footprint. The length, L, of the footprint

 0001-4966/2006/119(3)/1368/5/$22.50

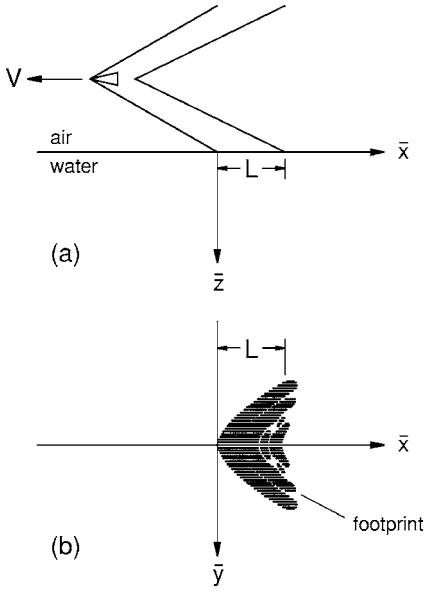


FIG. 1. (a) Elevation view of a body traveling supersonically in air with speed V. Intersections of the nose and recovery shock cones with the water surface define a surface footprint length, L. (b) View looking downward on water surface. The surface footprint is of finite extent in both the $\bar{x}$ and $\bar{y}$ directions.

[see Fig. 1(b)] can be defined as the distance along the flight track which encloses significant overpressures (say values which exceed 5% of the maximum in absolute value.)

The three-dimensional overpressure field, $\bar{p}(\bar{x},\bar{y},\bar{z},\bar{t})$, is assumed to satisfy the linear wave equation beneath the water surface ($\bar{z}>0$),

$$c^2\left(\frac{\partial^2}{\partial\bar{x}^2}+\frac{\partial^2}{\partial\bar{y}^2}+\frac{\partial^2}{\partial\bar{z}^2}\right)\bar{p}=\frac{\partial^2}{\partial\bar{t}^2}\bar{p}. \tag{1}$$

At the surface ($\bar{z}=0$), the following time-dependent boundary condition is satisfied

$$\bar{p}(\bar{x},\bar{y},0,\bar{t})=\begin{cases}P_{\max}f(\bar{x}_a,\bar{y}) & \text{for } (\bar{x}_a,\bar{y})\in R\\ 0 & \text{for } (\bar{x}_a,\bar{y})\notin R\end{cases}, \tag{2}$$

where

$$\bar{x}_a=\bar{x}+V\bar{t}$$

is the time-advanced position along the flight direction, and $f(\bar{x}_a,\bar{y})$ is an arbitrary function which describes the distribution of overpressure amplitude within the surface footprint. This shape function has a maximum value of unity, and $P_{\max}$ denotes the maximum overpressure within the footprint. In (2), R has been used to denote the footprint region, i.e., the set of points on the surface where the overpressure differs significantly from zero.

By defining dimensionless spatial coordinates as

$$x=\bar{x}/L,\quad y=\bar{y}/L,\quad z=\bar{z}/L \tag{3}$$

and dimensionless time and overpressure as

$$t=\bar{t}V/L,\quad p=\bar{p}/P_{\max} \tag{4}$$

the wave equation and boundary condition may be cast into nondimensional form

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)p=\frac{V^2}{c^2}\frac{\partial^2}{\partial t^2}p, \tag{5}$$

$$p(x,y,0,t)=\begin{cases}f(\xi,y) & \text{for } (\xi,y)\in R\\ 0 & \text{for } (\xi,y)\notin R\end{cases}. \tag{6}$$

It is clear from the dimensionless boundary condition (6) that the solution, $p(x,y,z,t)$, will depend only on the dimensionless time-advanced in-track coordinate

$$\xi=x+t \tag{7}$$

the dimensionless cross-track coordinate, y, and the dimensionless depth, z. Thus, the governing equations reduce to the following simple form:

$$\left(\beta^2\frac{\partial^2}{\partial\xi^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)p(\xi,y,z)=0, \tag{8}$$

$$p(\xi,y,0)=\begin{cases}f(\xi,y) & \text{for } (\xi,y)\in R\\ 0 & \text{for } (\xi,y)\notin R\end{cases}, \tag{9}$$

where

$$\beta^2=1-V^2/c^2 \tag{10}$$

is a real, positive constant, since $V<c$. The solution is now obtained by using the two-dimensional Fourier transform, s, of the overpressure field, $p(\xi,y,z)$. Specifically, define

$$s(k,q,z)=\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}p(\xi,y,z)e^{-iqy}dy\right]e^{-ik\xi}d\xi, \tag{11}$$

where k and q represent the in-track and cross-track wave number, respectively. Applying this transform to the wave equation (8) gives

$$-(\beta^2k^2+q^2)s+\frac{\partial^2 s}{\partial^2 z}=0 \tag{12}$$

and applying the transform to the boundary condition (9) gives

$$s(k,q,0)=\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}f(\xi,y)e^{-iqy}dy\right]e^{-ik\xi}d\xi. \tag{13}$$

Following the approach of Sawyers,[1] separation of variables is used to solve the transformed wave equation (12). Specifically, assume the Fourier transform of the solution is of the following form:

$$s(k,q,z)=F(k,q)Z(z). \tag{14}$$

Substituting (14) into (12), and separating variables gives

$$\beta^2k^2+q^2=\frac{Z''}{Z}=\alpha^2, \tag{15}$$

where α^2 is a real, positive separation constant. Equation (15) first implies that

$$Z=A_0e^{-\alpha z}+B_0e^{\alpha z}, \tag{16}$$

where A_0 and B_0 are constants, with the necessity that $B_0=0$, since a physically realistic solution cannot grow with depth. Second, (15) implies that

$$\alpha = \sqrt{\beta^2 k^2 + q^2}, \tag{17}$$

which shows that the depth decay rate is a function of the weighted sum of the two wavenumbers squared. (In Sawyers' two-dimensional solution, the cross-track uniformity of the solution to $\pm\infty$ in y implies vanishing cross-track wave number, i.e., $q=0$.)

The transformed solution at depth is thus

$$s(k,q,z) = s(k,q,0)e^{-\alpha z}, \tag{18}$$

where $s(k,q,0)$ is given by (13) and α is given by (17). The final solution is now obtained by taking the two-dimensional inverse Fourier transform of s,

$$p(\xi,y,z) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} s(k,q,z)e^{iqy}dq\right]e^{ik\xi}dk. \tag{19}$$

III. SAMPLE CALCULATIONS

The theory may be implemented numerically by using the discrete Fourier transform. Three problems are now solved using 512 equally spaced points in both the x and y directions. The spatial interval between points is set to

$$\Delta x = \Delta y = 0.01$$

and the (x,y) domain is delimited by

$$-2.05 \leqslant x \leqslant 3.06, \quad -2.56 \leqslant y \leqslant 2.55.$$

These limits were chosen with the idea in mind that the surface footprint, if *finite* in both x and y, is roughly limited in extent to the region $0 \leqslant x \leqslant 1$ and $|y| \leqslant 0.5$. That is, the surface footprint (overpressure distribution) is defined with roughly 10 000 points (100 in both x and y.) In the analytic representation of the Fourier transform, an undisturbed border of zero overpressure, extending to $\pm\infty$ in both x and y, is assumed to surround this footprint. Of course, for numerical evaluation via discrete transforms, the domain of computation must be truncated at some finite distance from the footprint region. The domain limits shown earlier reflect the fact that an infinite domain was simulated by assigning zero overpressure amplitude at roughly 200 points surrounding the footprint in each direction. In summary, 512 points were used to define the overpressure distribution at the water surface ($z=0$). Of these, only about 10 000 points in the center are normally needed to prescribe a *finite* sonic boom footprint. The remaining points are used to simulate a surrounding border region of zero amplitude.

The first problem is for infinite aspect ratio ($AR=\infty$) and serves to test the method in comparison to Sawyers' analytic solution for an N wave which is uniform in the cross-track coordinate, y. To mimic the infinitely uniform analysis with the current method, the N wave at the water surface ($z=0$) was initialized in a strip which extended across all y values. Specifically, the boundary values ($z=0$) were set as follows:

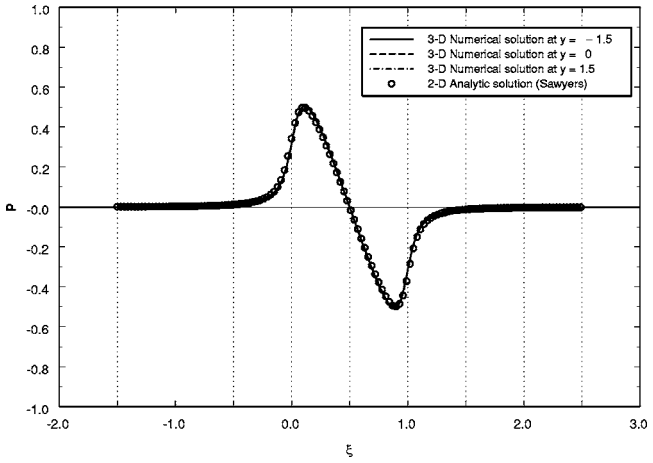


FIG. 2. Results for a surface N wave on $0 \leqslant \xi \leqslant 1$, which is uniform in y. The calculations are for dimensionless depth $z=0.1$ and speed ratio $V/c=9/16$. The three-dimensional results, using the current numerical procedure, are seen to be in excellent agreement with the two-dimensional analytic result given by Sawyers.

$$p(x,y,0) = \begin{cases} 1-2x & \text{for } 0 \leqslant x \leqslant 1 \text{ and } -2.56 \leqslant y \leqslant 2.55 \\ 0 & \text{otherwise} \end{cases}. \tag{20}$$

Since the edges of the computational grid correspond to $\pm\infty$, the footprint for this test case corresponds to $AR=\infty$. Results are shown in Fig. 2 for depth $z=0.1$ and for speed ratio $V/c=9/16$. (This speed ratio corresponds to a flight Mach number of roughly 2.5.) The overpressure distribution is seen to be uniform in y, as expected. The three-dimensional numerical results (plotted at three different values of y) are seen to be in excellent agreement with Sawyers' two-dimensional analytic result.

The second problem illustrates cross-track falloff of overpressure amplitude, when the surface footprint is of *finite* extent in the y direction. For this test, the N wave on the surface was enveloped by $\cos \pi y$, with the boundary values set to

$$p(x,y,0) = \begin{cases} (1-2x)\cos \pi y & \text{for } 0 \leqslant x \leqslant 1 \text{ and } |y| \leqslant 0.5 \\ 0 & \text{otherwise} \end{cases}. \tag{21}$$

Results at depth $z=0.1$ and for speed ratio $V/c=9/16$ are shown in Fig. 3. A decrease of overpressure amplitude with increasing $|y|$ is observed. It is also observed that the underwater amplitude directly beneath the flight track ($y=0$) is smaller in this case than the corresponding $y=0$ disturbance for the footprint of infinite cross-track extent. This is an expected result, due to a diminished contribution from outboard sources. (For reference, Table I lists the peak amplitudes of the four signatures plotted in Fig. 3.)

A third and final problem is now solved which represents an actual practical application of the current method. The previous two problems considered footprints which were very simple in shape, thus permitting simple mathematical

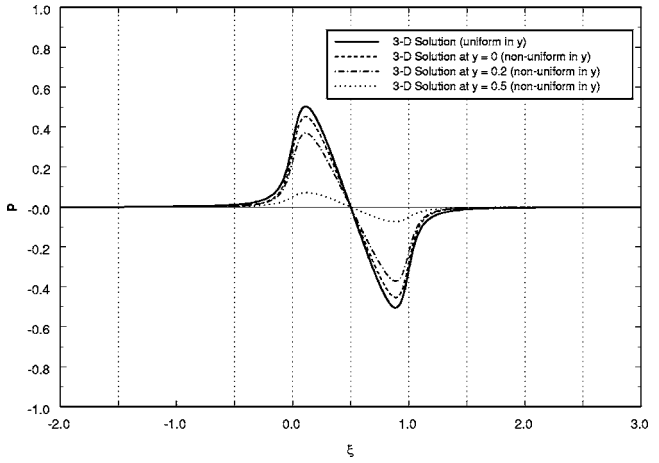


FIG. 3. Results for a surface N wave on $0 \leqslant \xi \leqslant 1$, which varies as $\cos(\pi y)$ for $|y| \leqslant 0.5$ and vanishes for $|y| > 0.5$. The calculations are for dimensionless depth $z=0.1$ and speed ratio $V/c=9/16$. As expected, the amplitude decreases as $|y|$ increases. It is also observed that the underwater amplitude at $y=0$ in the case of a surface disturbance of finite cross-track extent is smaller than the corresponding $y=0$ disturbance for the case of infinite cross-track extent.

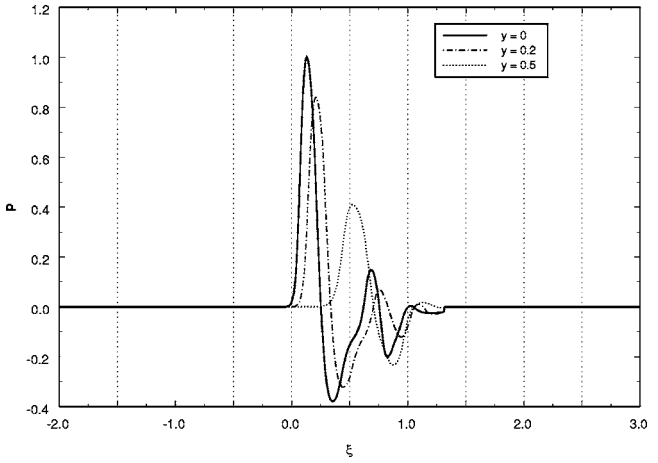


FIG. 4. Distribution of scaled overpressure on the ocean surface produced by a low-flying supersonic cruise missile. Variation along the direction of flight is shown for three different span-wise locations. The missile is 21 ft in length and 28 in. in diameter, flying straight and level at an altitude of 20 ft and a constant speed of 2500 ft/s. Unit dimensionless distance on the abscissa corresponds to 45.625 ft in actual physical measure, while a scaled overpressure of unity on the plot ordinate corresponds to 2.92 psi in physical measure.

descriptions. And although of use in illustrating the three-dimensional method, these footprints do not correspond to realistic distributions of surface overpressure generated by the supersonic flight of any actual body. In the final problem, use is made of the surface overpressure distribution generated by a supersonic cruise missile (21 ft in length and 28 in. in diameter) flying straight and level at an altitude of 20 ft above the ocean surface and a constant speed of 2500 ft/s. The speed of sound in the underlying seawater is taken to be 4888 ft/s. The surface overpressure distribution was generated via a three-dimensional computational fluid mechanics simulation provided by Wang.[5] The shape of the footprint is shown in Fig. 1(b). The variation of surface overpressure with distance along the flight track is plotted at three cross-track distances in Fig. 4, while Figs. 5 and 6 show the underwater pressure distribution at depths $z=0.1$ and $z=0.2$, respectively. A dimensionless length of unity along the in-track direction (abscissa of the plots) corresponds to 45.625 ft in actual physical measure. And a scaled overpressure (ordinate of the plots) of unity corresponds to 2.92 psi in actual physical measure.

IV. SUMMARY AND CONCLUSIONS

The Sawyers theory for underwater noise due to sonic boom has been generalized to three dimensions. The three-dimensional theory can be readily implemented numerically. Three numerical cases were run on a 512 by 512-point grid, with execution taking less than 2 s per case on a standard desk-top personal computer. One of the test cases showed that the method correctly reproduces Sawyers' analytic result in the case of a disturbance which is uniform in cross-track direction. A second case gave quantitative results of how, for a simply shaped surface footprint, the overpressure amplitude falls off with cross-track distance. The final problem provided underwater sound predictions for an actual cruise missile flying at an altitude of 20 ft and a speed of 2500 ft/s.

It is noted that a separate, and quite different, approach to three-dimensional solutions of this problem has recently been given by Wang.[6] In Wang's method, the entire volume of water of interest is gridded in three dimensions (x,y,z). Standard finite-difference methods for elliptic differential equations can be used to obtain the solution. The Fourier transform method given here may have some advantage in that one need only compute the transform of the boundary

TABLE I. Peak amplitudes at the various cross-track positions shown in Fig. 3.

Cross-track variation	y	Peak amplitude
Uniform	⋯	0.504
Nonuniform	0	0.453
Nonuniform	0.2	0.370
Nonuniform	0.5	0.072

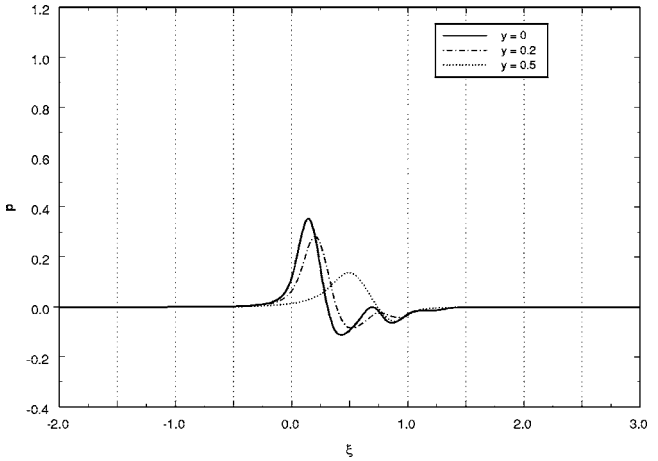


FIG. 5. Underwater distribution of scaled overpressure generated by a low-flying supersonic cruise missile. The results are for speed ratio $V/c=2500/4888$ and dimensionless depth $z=0.1$, which corresponds to a physical depth of 4.6 ft.

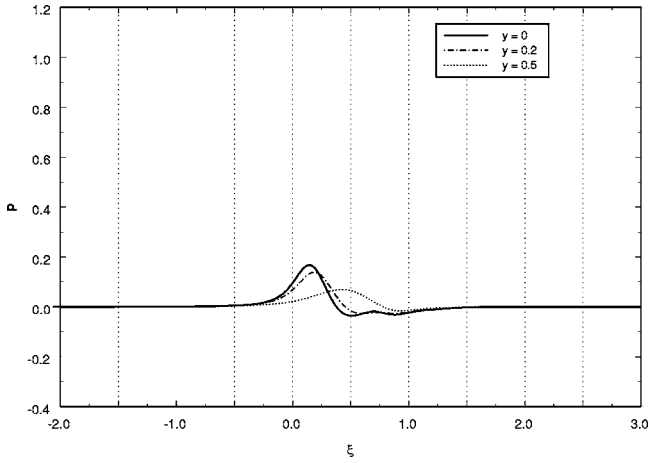


FIG. 6. Underwater distribution of scaled overpressure generated by a low-flying supersonic cruise missile. The results are for speed ratio $V/c = 2500/4888$ and dimensionless depth $z=0.2$, which corresponds to a physical depth of 9.2 ft.

condition in two dimensions $(x,y,0)$. A solution at any arbitrary depth may then be subsequently found by multiplying the surface transform by the depth-decay function and taking the inverse transform. Thus, there is no need for information at all depths to be simultaneously computed and stored.

The utility of the theoretical extension given in this paper lies mainly in predicting underwater noise levels resulting from low-flying supersonic aircraft and missiles.

ACKNOWLEDGMENTS

The author thanks John Edwards of the Space and Missile Center for supporting this work.

Nomenclature

Dimensional Variables

c = sound speed in water
L = in-track length of sonic boom signature at water surface
$\bar{p}$ = three-dimensional overpressure field within the water
$P_{\max}$ = maximum overpressure of sonic boom signature at water surface
$\bar{t}$ = time
V = horizontal flight speed
$\bar{x}$ = in-track horizontal coordinate along water surface
$\bar{y}$ = cross-track horizontal coordinate along water surface
$\bar{z}$ = vertical coordinate, measured positive downward from water surface

Dimensionless Variables

f = distribution of overpressure at water surface
i = $\sqrt{-1}$, unit measure along the imaginary axis
k = in-track dimensionless wave number
N = number of points in each coordinate direction of discrete two-dimensional Fourier transforms
p = scaled three-dimensional overpressure field within the water
s = two-dimensional Fourier transform of the scaled three-dimensional overpressure field within the water
q = cross-track dimensionless wave number
t = scaled time
x,y,z = scaled coordinates corresponding to $\bar{x},\bar{y},\bar{z}$, respectively
α = depth decay function
β = $\sqrt{1-V^2/c^2}$
ξ = $x+t$, in-track characteristic coordinate

[1]K. N. Sawyers, "Underwater sound pressure from sonic booms," J. Acoust. Soc. Am. **44**, 523–524 (1968).
[2]P. F. Intrieri and G. N. Malcolm, "Ballistic range investigation of sonic-boom overpressures in water," AIAA J. **11**, 510–516 (1973).
[3]V. W. Sparrow, "The effect of supersonic aircraft speed on the penetration of sonic boom noise into the ocean," J. Acoust. Soc. Am. **97**, 159–162 (1995).
[4]H. K. Cheng and C. J. Lee, "Sonic-boom noise penetration under a wavy ocean: Theory," J. Fluid Mech. **514**, 281–312 (2004).
[5]J. C. T. Wang, D. M. Moody, C. P. Griffice, J. R. Edwards, and A. A. Hashad, "Sonic boom and underwater sound pressure induced by low-altitude ramjet cruise missiles," AIAA paper-2006–0413, 44th AIAA Aerospace Science Meeting and Exhibit (January 2006).
[6]J. C. T. Wang, "Semi-similar formulation for three-dimensional underwater sound pressure propagation," J. Acoust. Soc. Am. (in press).

Mapping seabed variability: Rapid surveying of coastal regions

Charles Holland
The Pennsylvania State University, Applied Research Laboratory, State College, Pennsylvania 16804

(Received 5 July 2005; revised 5 December 2005; accepted 5 December 2005)

Directional broadband reverberation can be exploited for rapidly probing seabed variability over large areas O (10^3-10^4) km^2; the data need not be calibrated. Seabed variability thus mapped inherently combines spatial variability in geoacoustic properties that control reflection and those that control scattering. It is shown that in some cases, the variability can be separated. The method is applied to measurements in the Straits of Sicily where the data indicate the existence of two provinces where the statistics of the geoparameters are distinct. Three attributes of each province are quantified: the province boundaries, measures of spatial variability and/or uncertainty, and the location of discrete scattering features (that may lead to sonar clutter). One province is quite large and exhibits remarkably little spatial variability. The variability is probed on lateral scales $\sim 200 \times 400$ m, yet the ratio of the seabed geoparameters varies less than 3 dB from 300–1600 Hz over ~ 1000 km^2. While this result seems contrary to prevailing notions of extreme seabed variability in shallow water, it indicates that certain kinds of geophysical (and concomitant geoacoustic) variability may be averaged in such a way that the seabed appears approximately homogeneous to long-range acoustic systems. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2161439]

PACS number(s): 43.30.Pc, 43.30.Gv [AIT] Pages: 1373–1387

I. INTRODUCTION

Much is known about the processes that shape coastal sediment distribution. Such processes span a large range of space and timescales, ranging from eustatic fluctuations which are more or less global in extent but can have a profound effect on sedimentary layering from hundreds of meters to submeter scales, to biologic reworking which might be very local and affect only the upper few tens of centimeters. Riverine depositional processes play a very important role in many coastal environments as do storms and currents. These processes (as well as many others) conspire to create a rich variety of sediment properties distributions both in vertical and horizontal dimensions. Traditional instruments for geologic/geophysical mapping, e.g., seismic reflection and coring have been (and continue to be) powerful observational tools for identifying these processes.

While much is understood about the geological physical processes and how they shape sediment variability from a geologic perspective, relatively little is known about the scales of variability that affect long-range sonar systems. As a first step, one might be tempted to utilize the traditional geologic/geophysical mapping tools and assume that the observed variability also influences sonar variability. However, the scales of variability that are captured with these tools may not be the same as the scales of variability important for long-range sonar. Scales of seabed variability important for sonars depend in part of course on the particular sonar system (in particular, its frequency band and detection range) as well as the performance prediction requirements. A requirement for the prediction of details of a received waveform, for example, would be more sensitive to seabed variability than the prediction of the range-averaged response. In terms of seabed databases that support sonar performance prediction, the latter is generally the goal.

There are a few generalities that can be stated about relative sensitivities. For example, long-range sonar is sensitive to variability over a much smaller depth interval than a seismic reflection system at the same frequency. This is simply because the acoustic impedance of the seabed is much lower (and hence penetration is much greater) for a seismic reflection system at normal incidence than for a long-range sonar which operates at low grazing angles (where the acoustic field is often constrained to a few wavelengths below the water-sediment interface). Another aspect in which a long-range sonar might be expected to exhibit less sensitivity to seabed variability is for a range-averaged field quantity, where the acoustic field is averaged over scales of a cycle distance which may be of order 10^2-10^3 m (depending on water depth and sound speed profile). This means that some sediment variability at scales smaller than the cycle distance may not be important. As a counter example, some geologic mapping techniques tend to smooth out variability that can be important for sonar systems. One example in this category would be small outcrops, which might be either missed by a seismic reflection survey, or "smoothed out" because of the wide beamwidths of many geophysical sources.

Another tool to measure seabed variability could be the sonar system itself (or a surrogate). Here, we explore the potential for exploiting directional low-frequency reverberation and address the questions: "What kinds of variability can it detect and how well can that variability be measured?" Using reverberation data is an attractive prospect, because the data are relatively easy to collect and have very large coverage rates per unit time (as a gross comparison, roughly three orders magnitude faster than traditional geological/geophysical methods).

Ideally, we would like to obtain both the seabed properties and their variability using such data. However, we shall see in the following that there are severe uniqueness prob-

lems, i.e., it is very difficult (and in some cases impossible) to use reverberation data by itself to resolve the seabed properties. Nevertheless, we show that it is possible to use reverberation data to measure some kinds of variability. Thus, we view reverberation data as being commensurate with data from a seismic reflection system. That is to say that seismic reflection data do not provide information about seabed properties (because extracting density, velocity, and attenuation from such data has severe uniqueness problems); nevertheless, it is a most useful tool for studying the (combined) variability of these properties.

Though there is a substantial literature on the subject of using reverberation to extract seabed properties (see, e.g., Refs. 1–4 not intended to be exhaustive), this paper argues that reverberation data by itself can only be used to study variability (or changes in seabed properties) and generally not the properties themselves. The goal of the study was to determine the limits of using reverberation to study seabed variability, in other words to determine what kinds of variability can and can not be probed.

The approach draws on a range-averaged reverberation model developed by Harrison.[5] We begin by briefly exploring the sensitivities of the reverberation to seabed parameters (Sec. II). In order to examine the potential for measuring variability, we define a parameter called the scaled reverberation ratio, Γ and explore how variability (range and azimuth dependence) in the seabed parameters is manifest in the reverberation (hence in Γ). Of particular importance is the issue of detectability of seabed parameter variability in the reverberation data. A second key issue is that of uniqueness; in other words, if the seabed variability is detectable in the reverberation, can it be recovered unambiguously or is the variability coupled with variability in other parameters. The approach in Sec. II is demonstrated by applying it to measured reverberation data (Sec. IV). The experimental procedures are described in Sec. III. The data indicate the existence of several regions (or provinces) where the seabed parameters have distinct statistics within each region. The implications of the approach and the implications of the variability observed in the data set are explored in the final section.

II. DETECTING SEABED VARIABILITY: APPROACH

The potential for probing seabed variability is explored in this section, beginning with an examination of the sensitivity of reverberation to the seabed parameters followed by a description of the approach for obtaining seabed variability from reverberation data. Since the variability in seabed scattering and reflection are generally coupled in the reverberation, ways of decoupling the variability are also explored. The assumptions and limitations inherent in the approach are also described.

A. Relationship between reverberation and seabed properties

A variety of formulae have been developed for range-averaged waveguide propagation and reverberation using ray invariants and acoustic flux (e.g., Refs. 1 and 5). In general, the reverberation from a Pekeris waveguide can be written as

$$I = I_o \exp(-4\beta r)\frac{N}{H^2 r}\frac{c\tau}{2}\int_0^{\theta_c}\int_0^{\theta_c}\int_{\psi-\gamma/2}^{\psi+\gamma/2} M(\theta_i,\theta_o,\phi) \times \exp(-\alpha r\theta_i^2/2H)\exp(-\alpha r\theta_o^2/2H)\,d\theta_i\,d\theta_o\,d\phi, \quad (1)$$

where I_o is the source intensity, β is the attenuation in the seawater; r is range; N is the number of insonified patches (i.e., $N=2$ for an array of linear elements in a homogeneous environment except at endfire where $N=1$), ψ is the steering direction, γ is the horizontal beamwidth, c is the sound speed, τ is pulse length, H is water depth, θ_c is the critical angle, and M is the scattering kernel which depends upon incoming and outgoing vertical angles, θ_i, θ_o, and azimuthal angle ϕ. Finally, α is the slope of the plane wave intensity reflection coefficient, R, where below the critical angle, $R \sim \exp(-\alpha\theta)$. Note that in Ref. 1, α is defined in terms of the pressure reflection coefficient and thus is a factor of 2 smaller than the definition used here. For a homogeneous sediment half-space (see Ref. 6):

$$\alpha = 4\frac{\rho_b}{\rho}\frac{\chi}{(1-\chi)^{3/2}}\frac{c_b a_b}{2\pi 1000\ 20\log_{10} e}; \quad \chi = \left(\frac{c}{c_b}\right)^2, \quad (2)$$

where ρ_b, c_b, and a_b are the density, sound speed (m/s), and attenuation (in dB/m/kHz) of the sediment. The units are nepers/radian.

From Eq. (1), we can see that in the Pekeris waveguide, reverberation is sensitive to three properties of the seabed, the scattering law M, the critical angle θ_c (or equivalently the sound speed ratio at the seabed), and the reflection parameter α (or equivalently the density, velocity, and attenuation). So, there is a minimum of four seabed parameters (M requires at least two parameters) that must be known in order to predict reverberation. Thus, to invert for the seabed properties from reverberation data is a severely underdetermined problem. That is, there are many more parameters than independent data. Given that we cannot uniquely obtain the seabed parameters from the reverberation data alone, we wish to determine if reverberation can be used to learn something about the variability of the seabed parameters.

B. The scattering kernel

The scattering kernel is generally not known *a priori*. A widely used empirical model, Lambert's law, is given by

$$M = \mu \sin\theta_i \sin\theta_o \quad (3)$$

which leads to reverberation of the form

$$I = I_o \exp(-4\beta r)\frac{N\mu}{\alpha^2 r^3}\frac{\gamma c\tau}{2}(1-\exp(-\alpha r\theta_c^2/2H))^2 \quad (4)$$

differing slightly from Ref. 5 by the factor N and the inclusion of seawater attenuation β, which has a measurable effect at the ranges and frequencies examined here. The sensitivity of the reverberation to the various geoacoustic parameters is shown in Figs. 1(a)–1(c). Each panel shows the effect of changing one seabed parameter while holding the others constant at their mean value. Water depth is 100 m. Note that the reverberation curves (especially beyond a few km) show a similar decay rate for each of the param-

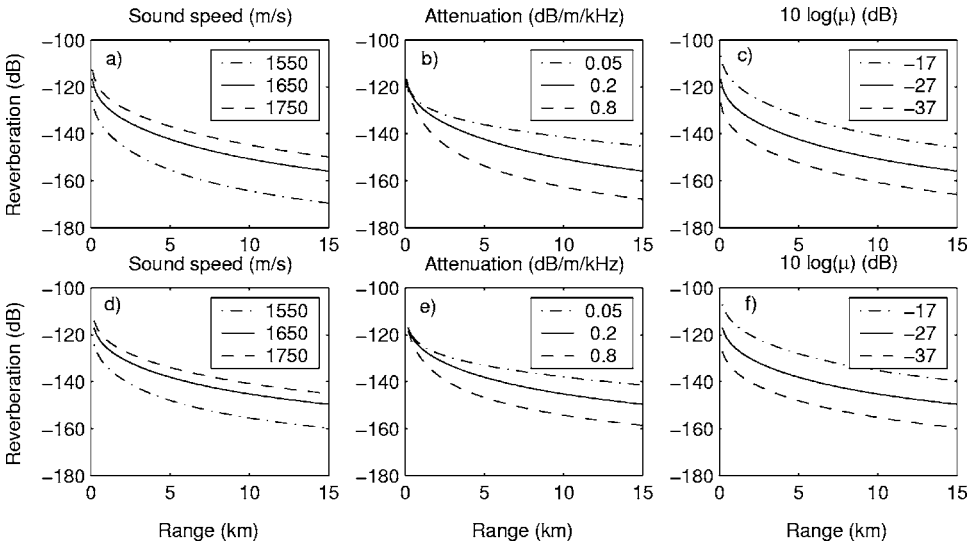


FIG. 1. Sensitivity of reverberation to various seabed parameters, (a)–(c) Lamberts law, (d)–(f) Lommel–Seeliger law.

eters. In other words, it would be difficult to uniquely invert the various seabed parameters using the decay rate of the reverberation. In this same regard, note that in Eq. (2) the parameter α is not a unique measure of the individual geoacoustic properties, i.e., many combinations of ρ_b, c_b, and a_b could give the same value of α_b.

In measurements of bottom scattering strength, the angular dependence is sometimes "flatter" than Lamberts law (e.g., Refs. 7 and 8) and better represented by the Lommel–Seeliger law

$$M = \mu \sin\theta_i \sin\theta_o/(\sin\theta_i + \sin\theta_o). \tag{5}$$

A comparison of the two scattering kernels are shown in Fig. 2. Many of the bottom scattering measurements made by the author seem to fit either a Lambert law a Lommel–Seeliger law, or somewhere in between. The reverberation with the Lommel–Seeliger law can be written as (see Ref. 16)

$$I \simeq I_o \exp(-4\beta r)\left(\frac{\pi}{2H}\right)^{1/2} \frac{\varepsilon N \mu}{\alpha^{3/2} r^{5/2}} \frac{\gamma c \tau}{2} \times (1 - \exp(-\alpha r \theta_c^2/2H))(\mathrm{erf}(\sqrt{\alpha r \theta_c^2/2H}))$$

$$\varepsilon = 1 - \log(\sqrt{2}+1)/\sqrt{2}. \tag{6}$$

Figures 1(d)–1(f) show the sensitivity of reverberation to the various geoacoustic parameters for the Lommel–Seeliger law scattering. The two important things to note are that: (1)

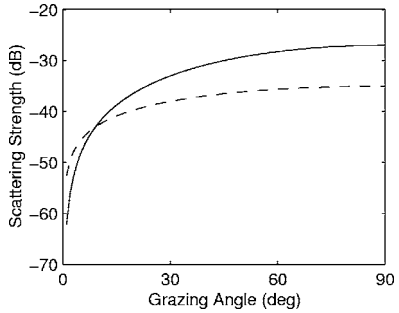


FIG. 2. Comparison of Lambert with $10\log\mu = -27$ dB (solid) and Lommel–Seeliger $10\log\mu = -32$ dB (dashed) scattering laws.

The reverberation curve especially beyond a few km has a similar rate of decay for each of the parameters and (2) that rate of decay with range is very similar to that of the Lambert scattering kernel (the difference is ~2 dB over 10 km). That suggests that in practice it would be difficult to determine the underlying scattering kernel (e.g., Lambert or Lommel–Seeliger) using solely long-range reverberation. This statement is supported, for example, in Ref. 4, where four different scattering kernels were fit to the same reverberation data with essentially the same fit statistics (see Fig. 4 Ref. 4).

Lambert's law has been used extensively and successfully in the modeling of observed reverberation. The insensitivity of reverberation to the scattering law, however, suggests that possibly many scattering laws could be made the fit the data equally as well. In other words, the relative success of Lambert's law may be simply due to the relative insensitivity of reverberation to the scattering kernel coupled with the predilection of the community to use it.

C. Scaled reverberation ratio

In our approach to studying seabed variability, we exploit this insensitivity as an advantage. That is, since we are not interested in the finding seabed parameters themselves, but only their variability, we can simply assume a fixed form of the scattering kernel. It is known[9] that scattering in much of this region is flatter than Lambert's law, i.e., falls off more slowly in angle than Eq. (3). However, we wish to demonstrate the method without any *a priori* knowledge and so Lamberts law will be assumed. Later on, we will show that this (incorrect) assumption adds only a modest uncertainty.

At long ranges, the final factor in Eq. (4) is ~1 in the long-range limit (see Sec. II E 1) and all the seabed properties can be grouped in a single factor

$$\frac{\mu}{\alpha^2} = \frac{I}{I_o}\frac{2r^3}{\gamma c \tau}\frac{1}{N}\exp(4\beta r). \tag{7}$$

If the seabed properties are allowed to vary in azimuth (ϕ), then in this long-range limit, changes in seabed properties could be observed by taking the scaled ratio of the two reverberation estimates. We define a quantity, Γ, as this scaled

reverberation ratio, which for a Pekeris waveguide and Lamberts law is:

$$\Gamma = \frac{\mu_2 \alpha_1^2}{\mu_1 \alpha_2^2} = \frac{I(r_2,\phi_2)}{I(r_1,\phi_1)} \frac{I_{o1}(\phi_1)}{I_{o2}(\phi_2)} \frac{r_2^3}{r_1^3} \frac{\gamma_1 \tau_1}{\gamma_2 \tau_2} \frac{N_1}{N_2} \times \exp(-4\beta(r_1 - r_2)). \quad (8)$$

For the Lommel–Seeliger law, the scaled reverberation ratio would be

$$\Gamma_{LS} = \frac{\mu_2 \alpha_1^{3/2}}{\mu_1 \alpha_2^{3/2}} = \frac{I(r_2,\phi_2)}{I(r_1,\phi_1)} \frac{I_{o1}(\phi_1)}{I_{o2}(\phi_2)} \frac{r_2^{5/2}}{r_1^{5/2}} \frac{\gamma_1 \tau_1}{\gamma_2 \tau_2} \frac{N_1}{N_2} \times \exp(-4\beta(r_1 - r_2)). \quad (9)$$

Equations (8) and (9) show that, in principle, it is possible to measure changes in the seabed properties between two patches on the seabed by taking a scaled ratio of the reverberation at those patches. Since reverberation can be measured rapidly, this is an attractive possibility. Furthermore, seabed changes can be detected without having a calibrated source or a calibrated receiver since only the ratio of the two transmissions and receptions are required. If the source is not omnidirectional in azimuth, a relative beam pattern is required. Also, note that if $N=2$, the ratio μ/α^2 or $\mu/\alpha^{3/2}$ is the mean value from the two beams.

For generality, we are interested in expressions that treat range dependencies in the environment including range-dependent scattering strength μ, reflection slope α, and bathymetry, H. In addition, we need to understand if the range dependencies can be detected unambiguously.

1. Range-dependent scattering strength, μ

Note that in Eqs. (8) and (9), the scattering strength μ is a "local" estimate, i.e., it depends only on the scattering strength in the illuminated patch. Therefore, Eqs. (8) and (9) implicity allow for range dependence in the scattering strength.

2. Range-dependent bathymetry, H

The case of range dependence in the bathymetry, (or case when the sound velocity profile is not isovelocity) is treated in Refs. 5 and 10, which have been benchmarked against a full-wave reverberation model.[11,12] For example, if water depth is slowly varying (commensurate with the adiabatic approximation), Γ for Lambert's law is

$$\Gamma = \frac{\mu_2 \alpha_1^2}{\mu_1 \alpha_2^2} = \frac{I(r_2,\phi_2)}{I(r_1,\phi_1)} \frac{I_{o1}(\phi_1)}{I_{o2}(\phi_2)} \frac{r_2^3}{r_1^3} \frac{\gamma_1 \tau_1}{\gamma_2 \tau_2} \frac{N_1}{N_2} \frac{H_{1s}^2 H_{2\,\mathrm{eff}}^2}{H_{2s}^2 H_{1\,\mathrm{eff}}^2} \times \exp(-4\beta(r_1 - r_2)), \quad (10)$$

where H_s and H_r are water depth at the source and receiver, respectively, and H_{eff} is the effective depth defined by Ref. 13 as

$$H_{\mathrm{eff}}(r) = H_s^2 H_r^2 / r \int_0^{r'} H^{-3}(r')dr'. \quad (11)$$

Strictly speaking, Eq. (10) is only valid when $N_{1,2}=1$; the more general case is discussed in Sec. II E 5.

3. Range-dependent reflection slope, α

The range dependence in the reflection slope $\alpha(r)$ can be treated by in an analogous way to H, by defining an effective reflection slope α_{eff}

$$\alpha_{\mathrm{eff}}(r) = \frac{1}{r} \int_0^r \alpha(r')dr', \quad (12)$$

so that in Eq. (1), α is replaced by the integral α_{eff}. For Lamberts law, the scaled reverberation is then

$$\Gamma = \frac{\mu_2 \alpha_{1\,\mathrm{eff}}^2}{\mu_1 \alpha_{2\,\mathrm{eff}}^2} = \frac{I(r_2,\phi_2)}{I(r_1,\phi_1)} \frac{I_{o1}(\phi_1)}{I_{o2}(\phi_2)} \frac{r_2^3}{r_1^3} \frac{\gamma_1 \tau_1}{\gamma_2 \tau_2} \frac{N_1}{N_2} \times \exp(-4\beta(r_1 - r_2)), \quad (13)$$

where in this case, $\Gamma(r)$ varies with range both due to μ and α. Note that the effect of the integral in Eq. (12) is to render α_{eff} a more slowly varying function of range than α. For example, if $\alpha(r)$ varies linearly with range, $\alpha(r)=\alpha'+\varepsilon r$ then the effective slope varies half as quickly or $\alpha_{\mathrm{eff}}(r) = \alpha' + \varepsilon r/2$. The case of $N=2$ represents the mean of the ratio $(\mu/\alpha_{\mathrm{eff}}^2)$ for the two ambiguous beams.

4. Detectability of range dependence in μ and α

It is important to consider how range variability in the seabed parameters μ and α is manifest in the reverberation data and in the scaled reverberation ratio. Scattering strength variability may occur and be detectable both at large scales (varying over scales of kilometers), as well as over small scales. The integration time and beam width determine the patch size and the lower bound of small-scale variability in μ that can be detected (for the data shown later, the smallest patch sizes are on the order of ~200 m by ~200 m). In the expressions above e.g., see Eqs. (10) and (13), rapid fluctuations in Γ are interpreted as fluctuations in μ, i.e., the data are assumed to be incoherent. The presence of coherence in the data would lead to overestimates of the variability of μ. That is to say that the fluctuations caused by coherence in the data may not be separable from the fluctuations caused by small-scale variability in the scattering strength. The data should be processed at a large enough bandwidth so that the data are commensurate with incoherent assumption.

Reflection slope, α, variability may occur both at large scales (varying over scales of kilometers), as well as over small scales. However, the variability is detectable only on large scales both because the range-averaged solution means that changes can only be sensed on scales of similar dimension or larger than the cycle distance and because the α_{eff} variability is less than the inherent α variability [see Eq. (12)].

The effects of large-scale variability in μ and α are probably indistinguishable in practice, at least without other independent data; that is to say we can only know about the ratio of the variability [e.g., Eq. (13)]. Furthermore, for a single observation of Γ, large-scale range dependence in either or both μ and α may be indistinguishable from uncertainty in the scattering law. Recall, that the rate of decay from different scattering laws (e.g., $r^{-5/2}$ and r^{-3}) is very difficult to distinguish in practice. If, for example, the large-

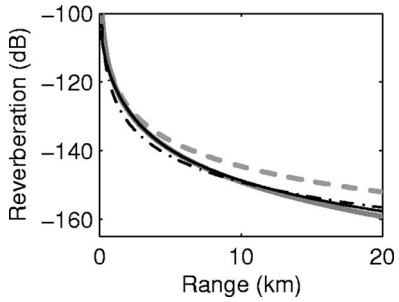


FIG. 3. Comparison showing effect of range-dependent α, range-independent $\alpha=0.5$ (dashed gray line); range-dependent $\alpha(r)=0.5+10^{-4}r$ where r is range in meters (solid gray line); approximate fit using the same scattering law and range-independent α (dash-dot black line); approximate fit using a different scattering law and range-independent α (solid black line). Given the fluctuations in reverberation, it may be difficult in practice to distinguish between the latter three parameter sets. In other words, a monotonic increase or decrease in the large scale α or μ may be difficult to recover unambiguously.

scale range dependence in μ/α^2 is monotonic with range, the rate of reverberation decay may not be distinguishable from uncertainty in the scattering law or even from the same scattering law and a single value for μ/α^2. An illustrative example of this is shown in Fig. 3, where reverberation along a given bearing with a range varying α may be indistinguishable from a range-independent α and a different scattering law or even the same scattering law with a range-independent α, given that fluctuations in the reverberation would usually mask such small differences. Thus, in practice, a range-dependent α along a single bearing is probably difficult to detect using reverberation.

5. Range-dependent H and α

For range dependence in both H and α, we have an integral of the following form:

$$K_{\text{eff}}(r) = H_s^2 H_r^2 / r \int_0^{r'} \alpha(r') H^{-3}(r') dr', \tag{14}$$

where K is a function of both H and α. Thus, in this general case, the scaled reverberation ratio cannot be defined with the seabed properties separable from the bathymetry. The approach taken here will be to assume that when there is significant range dependence in the bathymetry, the reflection slope α will assumed to be independent of range [using Eq. (10)]. When range dependence in the bathymetry is weak, Eq. (13) is used. If the bathymetry is known, it is easy to determine whether or not the effects of the range dependence of the bathymetry will be significant by calculating the term $H_{1s}^2 H_{2\,\text{eff}}^2 / H_{2s}^2 H_{1\,\text{eff}}^2$ in Eq. (10). If this term is ~ 1 over all ranges, the effect of range-dependent bathymetry is not significant. If bathymetric range-dependence effects are significant and the reflection slope α is range dependent, then there will be an (unrecoverable) bias in the estimate of $\alpha_{\text{eff}}(r)$. As an example, the errors in $\alpha_{\text{eff}}(r)$ [chain-dashed line of Fig. 4(b)] relative to the exact solution (solid line) are shown for a particular range-dependent α and bathymetry [Fig. 4(a)]. In this case, the estimate is biased higher than the true value, but the nature of the bias is highly dependent upon the particular form of $H(r)$ and $\alpha(r)$. For this simulation, the scattering strength was assumed to be constant (or known).

6. Implications

Directional reverberation contains information on the variability of the seabed in bottom-limited environments. The scaled reverberation ratio, Γ, provides a quantitative means of probing variability in the seabed scattering strength and reflection slope. Variability in scattering strength μ is observable both in range (along a given beam) and in azimuth (from beam to beam) in environments where the bathymetry is both range independent and range dependent. Variability in reflection slope α is observable in azimuth for both range-independent and range-dependent bathymetry, but range-dependent α will have a bias when there is significant bathymetric range dependence.

Up until this point, we have considered the ability or inability to separate the effects of μ and α variability by considering only a single observation of Γ. In the following, we consider the potential for using multiple observations of Γ to separate the variability. One of the key effects that permits some separation of parameter variability is the observation that the scattering strength is a local estimate, i.e., it depends only on the scattering strength in the illuminated patch. By contrast, the reflection loss slope is a "range-averaged" estimate, i.e., averaged over the entire range between the source/receiver to the scattering patch.

D. Uncoupling the variability using geometry

Since seabed variability has both local- and range-averaged effects on the observed quantity, Γ, these effects might be separated by judicious choice of geometry. Several

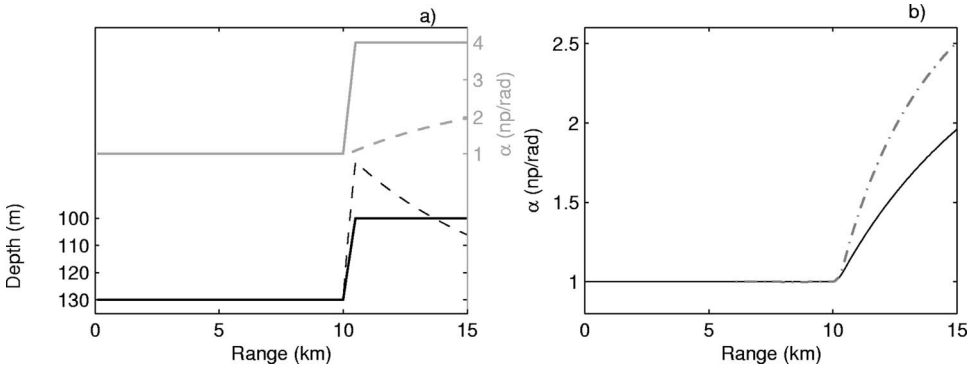


FIG. 4. An example of bias errors in estimating α_{eff} in the presence of range-dependent bathymetry (a) the environment for the simulation (solid) and H_{eff} and α_{eff} (dashed); (b) the exact (solid) and biased estimate of (α_{eff}).

categories of geometry where the effects of μ and α variability might be separable are considered in the following sections: (1) Case where the scattering patch is identical, (2) case where the scattering patch is different, and (3) case with radial symmetry.

1. Identical scattering patch

When the scattering patch location is identical between the two observations, the scattering kernel depends only on azimuth. For range-dependent bathymetry,

$$\frac{\mu(\phi_2)\alpha_1^2}{\mu(\phi_1)\alpha_2^2} = \frac{I_2(\rho)}{I_1(\rho)}\frac{I_{o1}(\phi_1)}{I_{o2}(\phi_2)}\frac{r_2^3}{r_1^3}\frac{\gamma_1\tau_1}{\gamma_2\tau_2}\frac{N_1}{N_2}\frac{H_{1s}^2 H_{2\,\mathrm{eff}}^2}{H_{2s}^2 H_{1\,\mathrm{eff}}^2} \times \exp(-4\beta(r_1 - r_2)). \quad (15)$$

In this and subsequent equations, α can in general imply a range-dependent α_{eff}. If there are multiple pings and the scattering patch is chosen so that r, ϕ are not linearly related, then the correlation between Γ and r, ϕ may help resolve the variability. For example, an insignificant correlation between Γ and azimuth would suggest that any variability is due to α not μ. A significant correlation between Γ and azimuth would be ambiguous, i.e., either α and/or μ could depend upon azimuth. In practice, seabed variability measurements for Eq. (15) are made on multiple pings on beams intersecting the patch of interest.

In the special case where the same patch is probed along the same bearing, Γ further simplifies to

$$\frac{\alpha_1^2}{\alpha_2^2} = \frac{I_2(\rho)}{I_1(\rho)}\frac{I_{o1}}{I_{o2}}\frac{r_2^3}{r_1^3}\frac{\gamma_1\tau_1}{\gamma_2\tau_2}\frac{N_1}{N_2}\frac{H_{1s}^2 H_{2\,\mathrm{eff}}^2}{H_{2s}^2 H_{1\,\mathrm{eff}}^2}\exp(-4\beta(r_1 - r_2)), \quad (16)$$

which is independent of scattering strength μ. So for this geometry, we can, in principle, separate the variability of α not μ. However, note that α_1 and α_2 are closely related, since they share a common range interval over $\min(r_1, r_2)$. Thus, α_2 will tend to change slowly since it cumulatively averages over a track [see Eq. (12)] and it may be difficult in practice to detect variability, except in cases where the change is very great between $\min(r_1, r_2)$ and $\max(r_1, r_2)$.

For a moving platform, seabed variability measurements using Eq. (16) can be made along the endfire beams on sequential pings. However, the endfire beams tend to be quite large and the forward endfire beam often is contaminated by towship noise.

2. Different scattering patch

When the scattering patch is different between the two observations, a special case of interest occurs when the two reverberation observations are co-linear. In this case, it is most convenient to make the observations along a given beam from a given pulse so that

$$\frac{\mu_2\alpha_1^2}{\mu_1\alpha_2^2} = \frac{I(r_2)}{I(r_1)}\frac{r_2^3}{r_1^3}\frac{N_1}{N_2}\frac{H_{1s}^2 H_{2\,\mathrm{eff}}^2}{H_{2s}^2 H_{1\,\mathrm{eff}}^2}\exp(-4\beta(r_1 - r_2)), \quad (17)$$

where the important simplification is that α_1 and α_2 are closely related, since they share a common range interval over $\min(r_1, r_2)$. Again, α will tend to change slowly since it is cumulatively averaged. In other words, for this geometry, range variability of μ would be easier to detect than range variability of α [as compared with the case of arbitrary geometry, e.g., Eq. (10)].

3. Radial symmetry

For a single ping in a monostatic geometry, the patch locations possess a radial symmetry. This symmetry can be exploited to help resolve some ambiguities in the observed variability. For example, consider a case where there is range independent bathymetry but α and/or μ are spatially variable. It was shown in the previous section that if the range dependence was monotonic, it could be difficult to separate this variability from uncertainty in other parameters, including the scattering kernel. Since it would be reasonably unlikely for the range dependence of α and/or μ to be radially symmetric around a given geographic position, a way to separate these effects would be to examine the azimuthal behavior of Γ. If there was a persistent variation with range/azimuth, it would suggest (all other effects being accounted for) that there was an underlying variation on α and/or μ. If, however, Γ was more or less isotropic, the simplest explanation would be that α and μ are independent of range and azimuth.

E. Assumptions and limitations

In this section, the assumptions and limitations of the method are assessed. One set of assumptions are associated with the model, these are that:

- The long-range limit is valid,
- Reverberation is dominated by seabed scattering (i.e., significant bottom interaction, negligible or predictable sea surface, and volume scattering),
- Adiabatic range dependence (technique is intended for continental shelf environments, where bathymetric slopes are generally modest),
- Existence of at least several propagating modes, and
- Weak focusing (e.g., see discussion in Ref. 10).

In addition, the following issues must be considered:

- Oceanographic variability,
- Source level variability,
- Left-right ambiguity for a linear towed array,
- Unknown scattering kernel, and
- Variability in multiple provinces.

These assumptions and issues are discussed in the following section.

1. Long-range limit

Using Eq. (4), the long-range limit can be defined in terms of ranges greater than n water depths:

$$n = \frac{r}{H} = \frac{-2\log(1-10^{-\varepsilon/20})}{\alpha\theta_c^2}, \tag{18}$$

where ε is the dB error between the long-range limit and the full expression. In order to obtain an approximate expression for n, we take the small angle approximation in Eq. (2) and picking a specific error, for example, $\varepsilon = 1$ dB, we have

$$n \cong 40\frac{\theta_c}{a_b}\frac{\rho c_b}{\rho_b c}. \tag{19}$$

Typically, attenuation a_b increases with increasing sound speed and critical angle, so (as reported in Ref. 5) n is often ~ 20 [note that Eq. (29) in Ref. 5 has different units of attenuation]. However, there may be instances where n is much larger than 20. One case would be where a_b is smaller than the Hamilton values used in Ref. 5, see, for example, Refs. 14 and 15. Another case would be more complex (e.g., layered) bottoms where the relationship between a_b and θ_c does not necessarily follow the empirical equations for a simple half-space. In terms of the long-range limit, the error term in Eq. (8) is

$$\frac{(1-\exp(-\alpha_1 r_1 \theta_{1c}^2/2H))^2}{(1-\exp(-\alpha_2 r_2 \theta_{2c}^2/2H))^2}, \tag{20}$$

and since both the numerator and denominator must be less than unity, the errors tend to cancel. However, in some cases, e.g., where the long-range limit is met for one observation and not for the other, the errors may not be negligible.

2. Sea surface scattering

As the sea surface roughness increases, two competing effects can occur, first that there is an increased reflection loss due to forward scattering. In the context of Eq. (2), we can write $\alpha = \alpha_b + \alpha_{sf}$ where the subscripts represent the bottom and surface contributions respectively. The second effect is an increased reverberation due to backscattering from the sea surface. These two effects and how they interact are not well understood by the community. A recent analysis,[4] however, showed that for a root-mean-square (*rms*) wave height of 0.35 m (wind speed of 9 knots) the reflection loss dominates. If the sea surface reflection loss was isotropic, the existence of $\alpha_{sf} > 0$ would act as a "noise floor" and limit the range of detectable α_b variability to values of $\sim > \alpha_{sf}$. For a rms wave height of 0.35 m, using the expression for the pressure reflection coefficient from Ref. 4, α_{sf} varies from about 0.024–1.5 nepers/rad from 100–1600 Hz. In such a case, variability in α_b much smaller than this would be difficult to detect.

3. Multiple propagating modes

The solutions derived from Eq. (1) above assume that there are many propagating modes. In the case where the water depth permits only one propagating mode, or mode stripping, has eliminated all but the lowest mode, the solution to Eq. (1) no longer contains α in the denominator. The transition to a single propagating mode was discussed in Ref. 5 and this transition occurs approximately at

$$\sqrt{\frac{H}{\alpha r}} \leqslant \frac{\lambda}{2H} \text{ or } \frac{4H^3}{\lambda^2 r} \leqslant \alpha, \tag{21}$$

where λ is the wavelength. Generally, α is not known *a priori*, but a reasonable estimate, based on a variety of locations is $\alpha \sim \kappa f$, where f is frequency in kHz and κ is roughly 2 rad^{-1} kHz^{-1} (e.g., see Ref. 1). Measurements of α in the Malta Plateau area[9] are within a factor of 2 of this rough estimate. For the frequencies and water depths of the experiment area (100–150 m), α is one or two orders of magnitude smaller than the requirement in Eq. (21). Thus, single mode effects are not expected to occur.

4. Variability in the oceanography and experimental variables

It is implicitly assumed that variability in propagation caused by (unknown) oceanographic variability is weak. In other words, we assume that the oceanography affects the two reverberation measurements in approximately the same way. This may not always be true. Thus, unknown range and time-dependent oceanography may corrupt the result, i.e., variations in transmission loss caused by oceanographic variability are explicitly assumed to be seabed variability.

The variability in Γ depends not only on seabed variability, but also on the uncertainties of the source level, range, and array heading (or equivalently beamwidth). Source level may dominate the uncertainty in instances where multiple transmissions are used; for the source used in this paper, the variability is expected to be a few dB.

5. Left-right ambiguity in towed arrays

For a towed array of omnidirectional hydrophones, there is a left-right ambiguity. In a range-independent environment [e.g., Eq. (8)], $N = 2$ for all beams except endfire. However, in a range-dependent environment, the bathymetry and/or μ/α^2 may vary between the beams. The total intensity in the long-range limit from the two beams (assuming Lamberts law) is

$$I = I_o \exp(-4\beta r)\frac{H_s^2}{r^3}\frac{\gamma c \tau}{2}\left(\frac{\mu^a}{(\alpha^\alpha H_{\text{eff}}^a)^2} + \frac{\mu^b}{(\alpha^b H_{\text{eff}}^b)^2}\right), \tag{22}$$

where superscripts a and b indicate the two ambiguous beams. For completely arbitrary range dependence, it is not possible to obtain the scaled reverberation ratio where the seabed properties are independent of the bathymetry. However, when water depth is slowly varying, a useful approximation is

$$\begin{aligned}\Gamma &= \left\langle \frac{\mu_2}{\alpha_2^2} \right\rangle \Big/ \left\langle \frac{\mu_1}{\alpha_1^2} \right\rangle \\ &\simeq \frac{I(r_2,\phi_2)}{I(r_1,\phi_1)}\frac{I_{o1}(\phi_1)}{I_{o2}(\phi_2)}\frac{r_2^3}{r_1^3}\frac{\gamma_1\tau_1}{\gamma_2\tau_2}\frac{H_{1s}^2 H_{2\,\text{eff}}^a H_{2\,\text{eff}}^b}{H_{2s}^2 H_{1\,\text{eff}}^a H_{1\,\text{eff}}^b} \\ &\quad \times \exp(-4\beta(r_1 - r_2)),\end{aligned} \tag{23}$$

where $\langle\ \rangle$ indicates the mean of beams a and b. If H_{eff} varies substantially from a given beam to its ambiguous side and μ and α are believed to be approximately constant, the

factor $H^a_{\text{eff}}\,H^b_{\text{eff}}$ would be replaced by $(H^a_{\text{eff}})^2+(H^b_{\text{eff}})^2$. In the case where one beam exhibits a significantly higher reverberation than the ambiguous beam at a given time, either because of propagation effects and/or a significant difference in scattering strength, the expression for Γ would be well approximated by Eq. (10) with $N=1$.

6. Errors in assumed scattering law

Since the scattering law is generally not known *a priori*, it must be assumed. For the data set in question, it is known through other analyses[9] that the underlying scattering is flatter than Lamberts law. If the scattering follows the Lommel–Seeliger law, the error in assuming Lamberts law in this case would be

$$\frac{\Gamma_{\text{Lambert}}}{\Gamma_{\text{L–S}}}=\left(\frac{r_1}{r_2}\right)^{1/2}\left(\frac{H_{1\,\text{eff}}}{H_{2\,\text{eff}}}\right)^{3/2}\left(\frac{H_{2s}}{H_{1s}}\right)^{2}. \tag{24}$$

In many environments, this error would be quite modest. For example, given a reference range r_1 at the geometric mean of the minimum and maximum ranges of interest, say 3 to 20 km, the maximum errors for range-independent bathymetry would be ±2 dB (at 3 and 20 km). This kind of error is probably not significant (or detectable) given that the fluctuations in the reverberation are often of that order or larger.

The fact that Γ is relatively insensitive to the choice of scattering law is not meant to suggest that knowledge of the scattering law is not important. Knowledge of the scattering law is quite important (and the bias errors can be significant) when attempting to estimate the scattering strength μ (e.g., see Ref. 16).

7. Mapping variability in multiple provinces

Consider the case of two distinctly different provinces (say A and B), where μ and α are more or less constant within each province. As discussed in Sec. II C, the reverberation (or Γ) is not very sensitive to changes in α, even when it rapidly changes across a province boundary. Thus, if source/receiver are in Province A, the reverberation from Province B will be controlled by α_{eff} which may be dominated by its value, α_A. This has several implications. First, there is limited ability to probe the sediment properties in a province outside of the province where the source/receiver is located. As an example, consider the environment in Fig. 4(a) but with range-independent bathymetry (to eliminate the separate issue of bias errors). For the source/receiver at 0 km, at ranges of a few km into the distinct province (ranges greater than 10 km), the value of is α_{eff} is nearly the same as the province in which the source/receiver were located ($\alpha_{\text{eff}}\sim 1$). The true local value of is $\alpha=4$, but this value cannot be sensed with this geometry, except at ranges of several times the distance from the source/receiver to the province boundary (assuming the province is large enough in spatial extent).

The second implication is that the averaging of reverberation (or Γ) at a particular patch from various source/receiver locations may lead to errors inasmuch as α_{eff} (and/or the bias errors discussed previously) at a particular geographic location is highly dependent on the source/receiver position. As an example, consider the environment in Fig. 4(a) with source/receiver at 0, 8, and 15 km and let us examine the reverberation at a patch at 12 km range. The α_{eff} from the three source/receiver positions (not considering bias errors from bathymetry) will be 1.4, 2.4, and 4 np/radian respectively; hence, the scaled reverberation ratios will differ relative to the 0 km range by 5 and 9 dB, respectively. Thus, averaging from multiple locations across province boundaries may lead to significant errors in the estimation of μ,α.

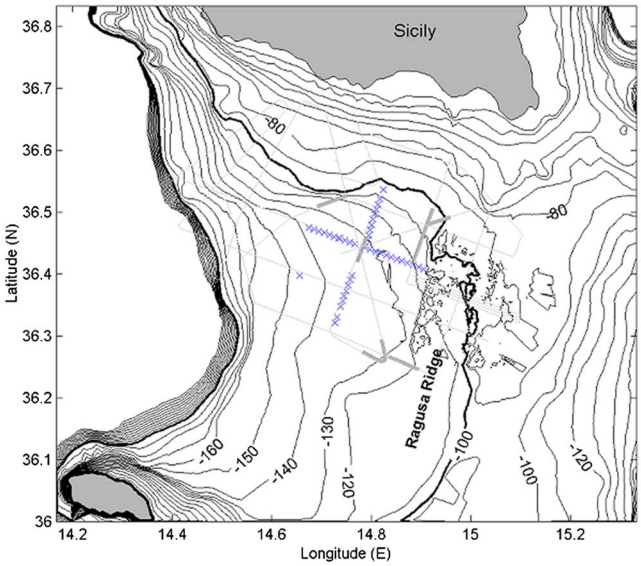


FIG. 5. Map of experimental area in the Straits of Sicily (Malta Plateau). Depths are in meters, and were taken from historical soundings, except for a small section along the Ragusa Ridge between about 36.3°–36.5° N. Shot locations are shown by "x." Seismic reflection survey shown in thin gray lines. Heavy gray lines are seismic reflection transects.

F. Sensitivities and provincing

The sensitivity of the seabed properties to the scaled reverberation ratio, Γ, in Eq. (10) is briefly examined. The sensitivity of Γ to μ is straightforward. Changes in μ (relative to the reference point) of a few dB or greater should be observable; values in the literature for μ in this frequency band, range from about -15 to -40 dB. The sensitivity of Γ to μ is identical between the Lambert and the Lommel–Seeliger laws.

The scaled reverberation ratio is also sensitive to α. For example, changes in α of a factor of 2 lead to a 6 dB change in the reverberation ratio for Lamberts law. For Lommel–Seeliger law, Γ is somewhat less sensitive to changes in α, e.g., changes in α of a factor of 2 would lead to a 4.5 dB change in the reverberation ratio. Recall that since α is cumulative along a given beam direction, it is easier to detect azimuthal (or beam to beam) changes in α than along a given beam.

The ability to measure seabed variability from reverberation data provides the opportunity to examine a concept used in seabed databases. Many seabed databases make the assumption that the seabed is laterally relatively "uniform" across some area, such an area is defined as a province. A

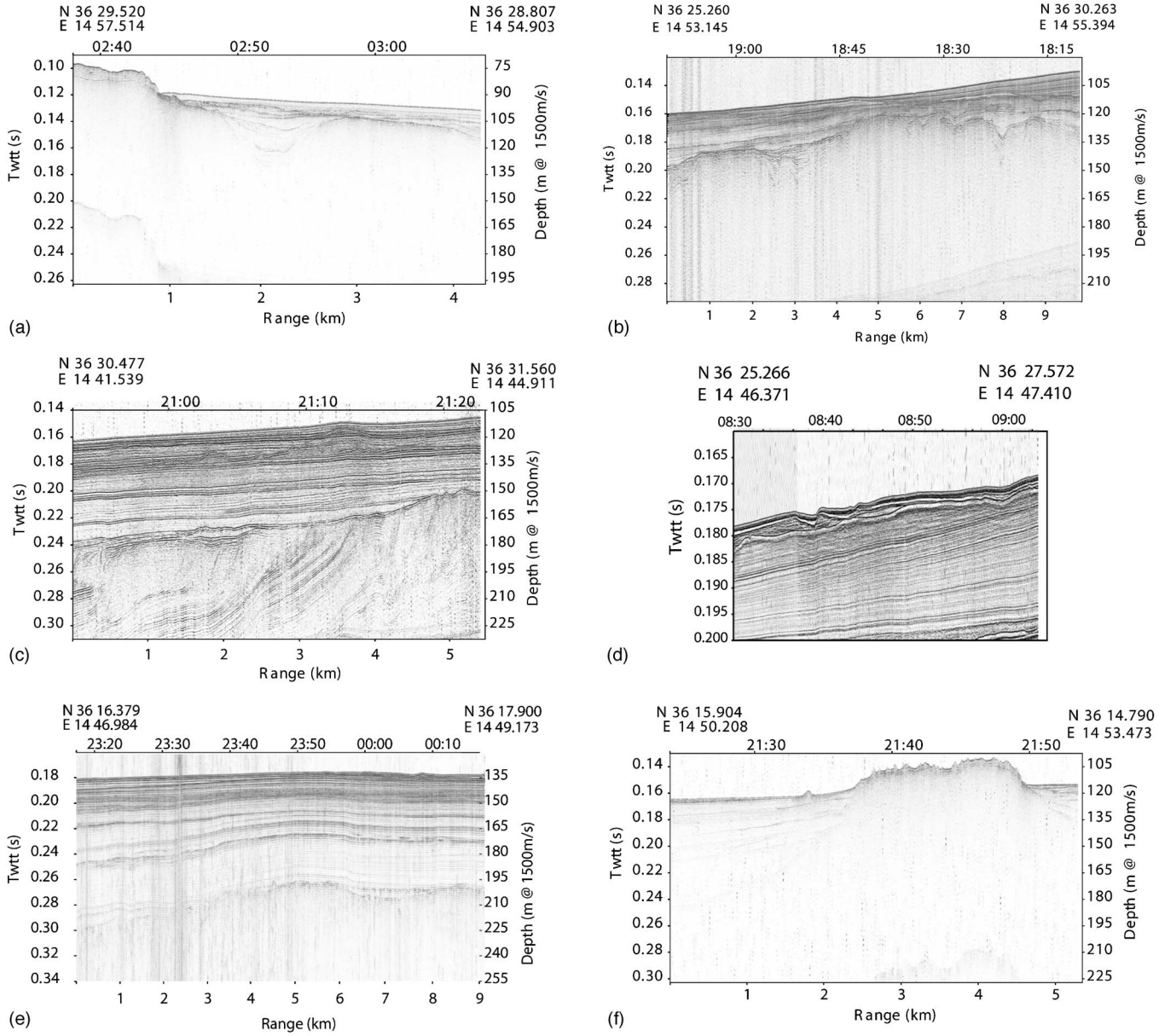


FIG. 6. Seismic reflection profiles showing stratigraphic variability in the area. Note that the data are presented at different amplitude scales (clipping levels) and vertical exaggerations. Locations of each transect are shown in Fig. 5 in the heavy gray lines, the order of the plots above (a)–(f), follows a counterclockwise direction starting from the northeastern-most transect in Fig. 5 northeastern-most transect in Fig. 5 showing ridge, (b) transect showing layering near ridge, (c) transect in the north-western transect, (d) transect near the center of the SUS pattern, core data show that the sediment lens a few tens of centimeters below the water-sediment interface has large cobbles up to 10 cm in diameter, (e) transect showing layering in the south-western corner, the large return at ~2.5 km is due to a wreck (f) transect across ridge in the southeastern part of the survey area.

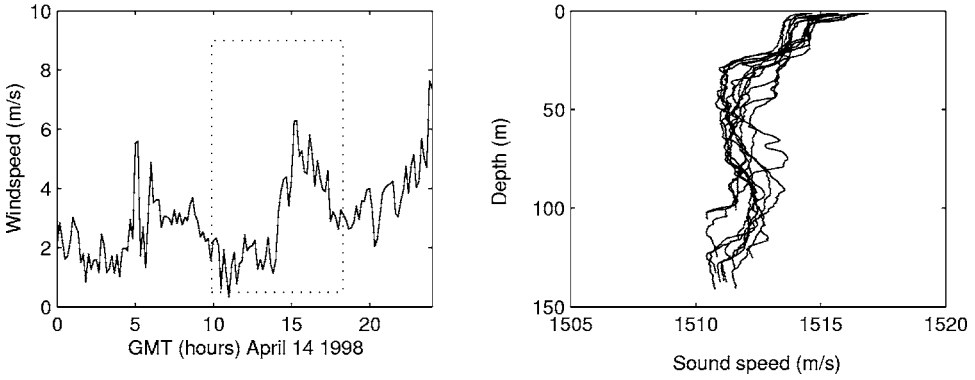


FIG. 7. Wind speed during the reverberation experiments (095200–181300 GMT) and 12 XBT measurements collected during the same period every ~50 min. The sound speed measurements include both time and space variability (i.e., were collected along the shot tracks shown in Fig. 5).

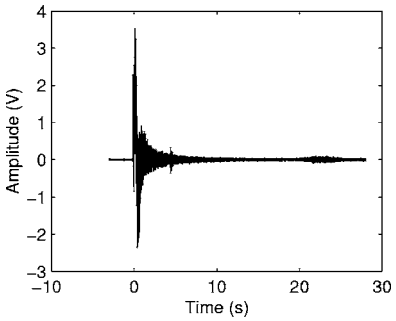


FIG. 8. Broadband beamformed reverberation time series from Ping 1 (southernmost ping) on the forward quartering beam. The data are referenced to the SUS initiation (time=0) and the direct arrivals (typically first few hundred milliseconds) are clipped.

purist may insist that the seabed is never uniform and that variability exists at all scales. Indeed, it is not difficult to find examples of closely spaced core samples that show significant variability. For sonar systems, however, this kind of variability may or may not be important.

III. EXPERIMENTAL DATA

Long-range reverberation data were acquired April 14, 1998 along two tracks on the Malta Plateau during the SCARAB98 experiment. The environment and the equipment employed are briefly described below.

A. The environment

The Malta Plateau (see Fig. 5) occupies the northern edge of the North African passive continental margin and is a submerged section of the Hyblean Plateau of mainland Sicily. A discussion of the geology of this area, along with seismic reflection data and core data can be found in Refs. 17–19. The region is divided by the Ragusa Ridge, which forms a spine ~15 km wide between Sicily and Malta. The ridge has exposed rock (presumably limestone) on its eastern and western edges. Between the exposed rock is a broad area of ponded sand (sometimes many tens of meters thick) with some rocky outcrops. West of the ridge, the seabed is blanketed with silty-clay sediment layer that is about 10 m thick at the 100 m depth contour and thins seaward; for example, at 128 m the layer is about 0.80 m thick. Below the silty clay, there is a thick (several hundred meters) sequence of unconsolidated sediments. The properties of the silty-clay layer and some of the deeper layers are given in Ref. 7. Seismic reflection data presented in Fig. 6 show some of the scales of variability present in this area as seen through this normal incidence system.

The wind speed during the experiment was generally less than 6 knots [see Fig. 7(a)]. Waverider buoy data, collected at a sampling frequency of 2.56 Hz, were averaged to provide the rms height over 30 min intervals. During the reverberation experiment, the mean rms height was 0.16 m with a standard deviation of 0.007 m; the dominant wave period was 8 s. Using the surface reflection formulation in Ref. 4, the slope of the surface reflection loss, α_{sf} varies 0.007–0.47 nepers/radian from 100–1600 Hz, respectively. This is far below the values of α_b expected in this area (see Ref. 9). Thus, the sea surface loss plays a negligible role in mapping seabed variability for this data set.

Sound speed profiles were collected using expendable bathythermographs (XBTs) every ~50 min during the experiment from the ship as it executed the reverberation pattern (see Fig. 5). The sound speed profiles [see Fig. 7(b)] are slightly downward refracting with a gradient of ~0.018 s^{-1} and exhibit only a modest variability in space and time.

B. Source and receiver

Impulsive sources were employed [Mk61 Sound Undersea Signal (SUS)] at 91 m depth approximately every 5 min along each track (see Fig. 5). The SUS were launched from the vessel and initiated within a few hundred meters of the receiver. In the analysis, the geometry is assumed to be monostatic, which is a reasonable approximation given that the ranges of interest are greater than 3 km.

The receiver was a three-aperture nested horizontal array of 256 elements with element spacings of 0.5, 1, and 2 m towed at a depth of ~50 m. The data were digitized at 6000 Hz sampling rate and low-pass filtered at 1780 Hz with a seven-pole six-zero elliptic (70 dB per octave rolloff) antialias filter. Data presented here were time-domain beamformed with Hanning shading on the 0.5 m aperture. An example of the beam-formed reverberation time series on one beam (the forward quartering beam) is shown in Fig. 8.

In the processing, an integration time of 0.25 s was employed, which corresponds to a radial patch dimension of 189 m. The data were incoherently averaged in frequency in 100 Hz bands from 100–1800 Hz. Background noise levels on each beam were estimated for a 2 s window before the direct blast and only reverberation levels greater than 6 dB above the background noise were used in the analysis. The range to a particular scattering patch, r, was estimated simply by $r=c_{\mathrm{av}}t/2$, where c_{av} was computed from the measured sound speed profile. As an example of the processing, the reverberation data at 800 Hz are shown in Fig. 9(a) and the corresponding scaled reverberation ratio in Fig. 9(b).

IV. RESULTS

The reverberation data were manipulated to examine azimuthal variability in two ways: With a fixed scattering patch [Eq. (15)] and with a variable scattering patch Eq. (10).

A. Fixed scattering patch

We begin by examining the behavior of the scaled reverberation ratio for a fixed patch location [see Eq. (15)]. Patches were randomly selected in various areas, including the area west of the ridge and on the ridge itself.

Figure 10 shows a representative result where the reference location (36° 17.4′N 14° 48.6′E) was placed west of the Ragusa Ridge. The location of each of the nine source pings is represented in Fig. 10(a) by a blue circle, the scattering patch is indicated by a red asterisk and the ambiguous beam by a green "x." In practice, the ambiguous beam is slightly smeared due to small changes in the array heading along a track. The smearing is plotted in Fig. 10(a) but is

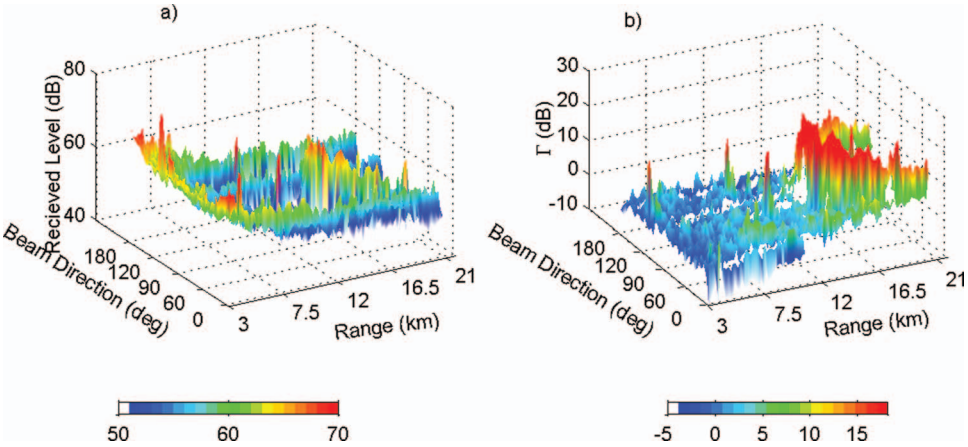


FIG. 9. 800 Hz reverberation from Ping 1(a) received level, and (b) scaled reverberation ratio Γ.

small enough (the patch centers are separated by less than 200 m) not to be noticeable on that scale. Since the ambiguous patch is essentially the same for each ping, the ambiguous beam does not add any uncertainty to this particular data set.

Fig. 10(b) shows the variation of the scaled reverberation as a function of frequency for the various pings. The reader is reminded that for reverberation data in the long-range limit, it is not possible to know anything separately about the scattering coefficient μ and reflection loss slope α, but only their ratio. For example, it is not known whether the minimum at about 600 Hz is controlled by μ or α or both. By normalizing the scaled reverberation by one of the observations (in this case, the first ping was taken), the scaled reverberation ratio Γ is computed [see Fig. 10(d)]. The standard deviation (computed in $\log_{10}$ space) is generally smaller than 2 dB, and increases just slightly with frequency [Fig. 10(e)]. The relationship between the variation in the scaled ratio and azimuth indicates no strong correlation [Fig. 10(f)].

The fact that the scaled reverberation ratio shows little variability over these pings, means that the scattering strength μ does not vary over the measured azimuths at that point and that α does not vary from one source-patch-receiver path to the other. The other explanation is that the variability in α and μ cancel each other out, which is possible, but less likely according to Occam's Razor. In other words, the seabed in the triangular region defined by the source-patch-receiver locations, roughly 50 km^2, appears to be homogeneous with respect to the reflection slope α. Various random points in this area (west of the ridge) all show similar results, i.e., the standard deviation of the scaled reverberation ratio is ~2 dB, with no evident dependence on azimuth.

Several reference patches were also selected on the Ragusa Ridge. Figure 11 shows the results one of the patches. The scaled reverberation of Fig. 11(b) is substantially higher than that in the basin. The standard deviation of the scaled ratio is still reasonably small, 3 dB or less. There is, however a clear trend in the variability, the variability increases as a function of increasing azimuth and frequency [see Fig. 11(f)]. Since the azimuthal variation of α appears to be small west of the Ridge, this variability is interpreted as azimuthal variability of the scattering strength. Thus, over a range of 35° in azimuth, the scattering strength appears to be independent of azimuth at a few hundred Hz but varies ~10 dB at 1600 Hz.

A clear example of where the seabed is not homogeneous (shown in Fig. 12) occurs when the track crosses the ridge boundary. Note the large variation in Γ. The first ping, which was used as a reference, clearly has a significantly different level than the others. But even in the other pings, there is considerable variability. In a blind experiment, it would not be possible from these data alone to determine the cause of the inhomogeneity, or where the province bound-

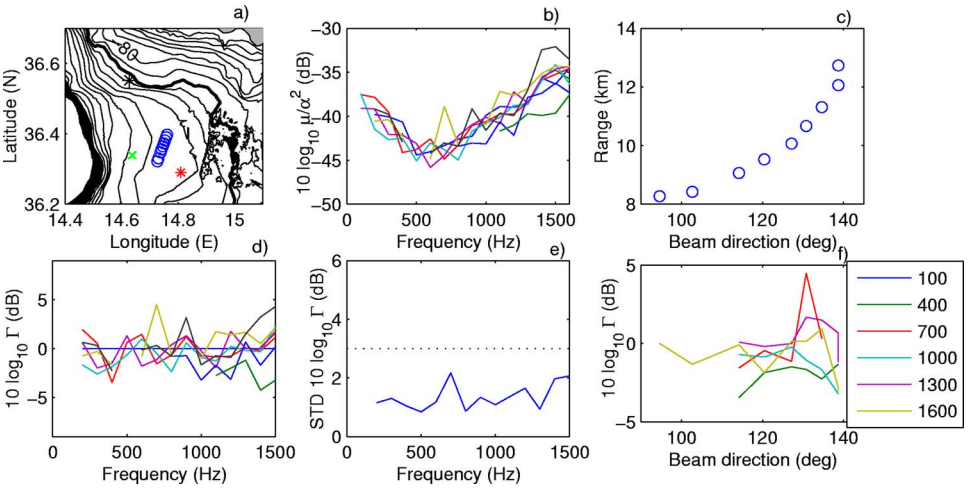


FIG. 10. (Color online) Reverberation processed at a single point west of the Ragusa ridge, (a) map of area with location of pings (○), the scatter location (*), and the ambiguous beam scatter location (×); (b) the results of nine pings processed using Eq. (15); (c) dependence of scattering patch on beam steering angle and range; (d) scaled reverberation ratio; (e) standard deviation of Γ; and (f) dependence of Γ on beam steer direction.

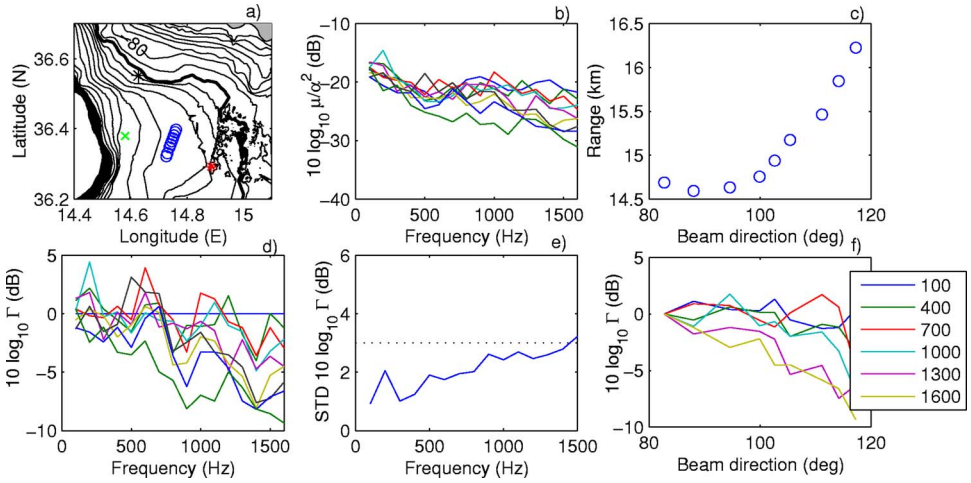


FIG. 11. (Color online) Reverberation at a single point on the western rim of the Ragusa ridge, various panels are explained in Fig. 10.

ary(ies) were that could explain such variability. Detecting and mapping province boundaries is explored in the following section.

B. Variability from radial symmetry

In order to identify seabed variability over a wide area, the scaled reverberation ratio Γ is computed for arbitrary geometries. In particular, consider the behavior of Γ for a single ping. Results at 800 Hz for the first ping in the track are shown in Fig. 13, where blank areas indicate a reverberation-to-noise ratio less than 6 dB. The reference point was taken at a random point west of the ridge [same location as in Fig. 10(a)].

There are two salient points in these data. First, that Γ shows modest variability ($\sim\pm 2$ dB) west of the ridge, consistent with results from the previous section. The presence of isolated "speckles" in Γ are indicative of local changes in scattering; the speckle cannot be caused by variability in α because along a given beam, α_{eff} must be slowly varying [see Eq. (12)]. The second salient point is that all along the Ragusa Ridge, Γ is much higher than two standard deviations ($\sim$4 dB). Data from other headings (not shown) establish that the high levels correspond to the Ragusa Ridge (and not from the ambiguous beam).

C. Province attributes

The reverberation data considered thus far indicate that this area can be divided into two distinct regions or provinces where the scaled reverberation ratio is similar. These two provinces will be designated "basin province" (the region to the west of the ridge) and the other "ridge province." For this study, the key attributes important for defining the provinces are: (1) The province boundaries, (2) a mean or median frequency-dependent seabed descriptor, (3) measures of spatial variability and/or uncertainty, and (4) discrete features.

Before discussing the province boundaries, statistics of the spatial variability are considered. Histograms of Γ were computed for the entire time series [Fig. 14(a)], as well as the two provinces [Figs. 14(b) and 14(c)]. The two provinces have distinctly different means and standard deviations (see Fig. 15) at all frequencies. The calculations for Γ have taken into account range-dependent bathymetry, as well as regions where the scattering level is dominated by a single beam. The difference between using range-dependent bathymetry over the ridge and range-independent bathymetry (if none were available) is $\sim$3–4 dB.

The range-dependent modeling at the ridge boundary may be in error if the bathymetric slopes are larger than the adiabatic approximation permits. The seismic reflection data

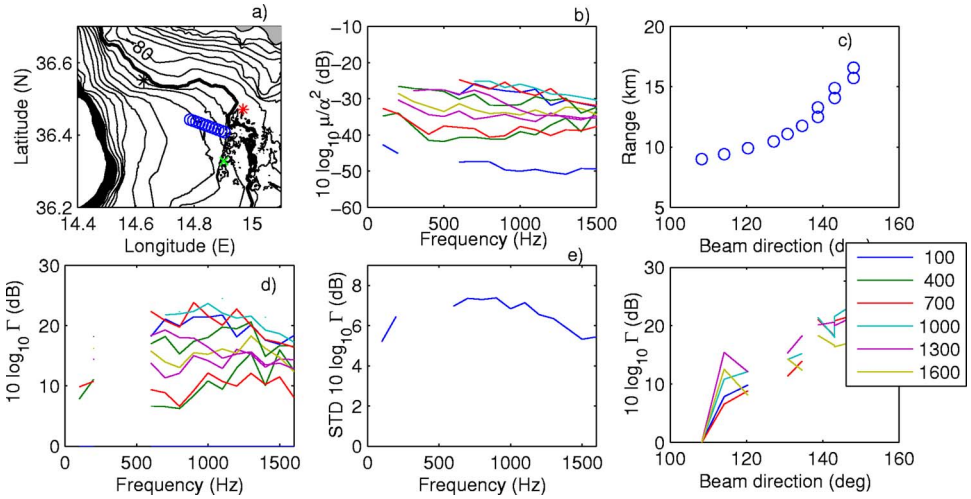


FIG. 12. (Color online) Reverberation at a single point on the northwestern rim of the Ragusa ridge, various panels same as Fig. 10.

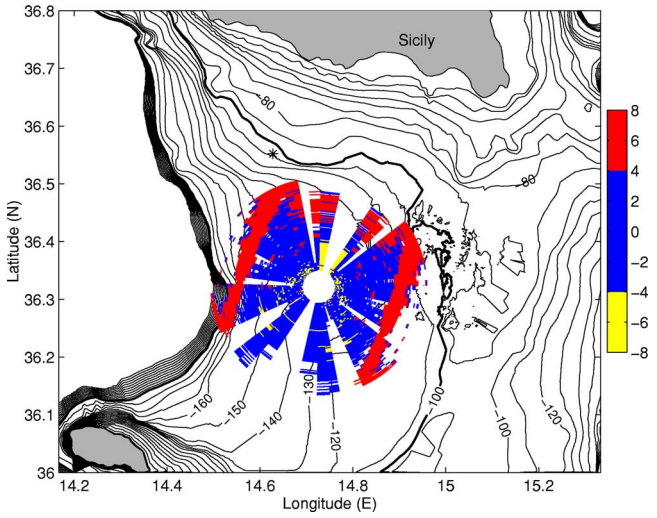


FIG. 13. (Color online) Scaled reverberation ratio Γ (in dB) at 800 Hz for the first ping in the south-north track.

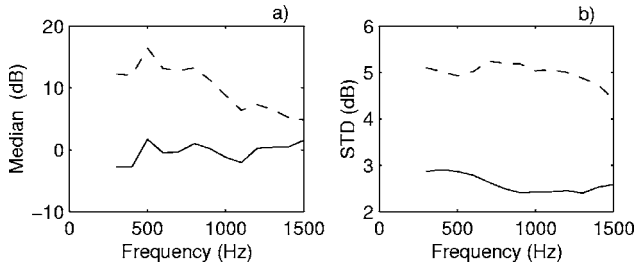


FIG. 15. Frequency dependence of single ping statistics for the basin (solid) and ridge provinces (dashed); (a) median and (b) standard deviation. The beam closest to endfire was not included in the statistics.

indicate that the slopes on at the ridge boundary are about 5°, however, the wide beamwidth of the source (Uniboomer) means that if steeper slopes existed, they may not be detectable with that system. If the adiabatic approximation is not valid, the actual values for Γ on the ridge would be somewhat smaller, i.e., the distribution [Fig. 14(c)] would contain fewer very large values of Γ. Range-dependent calculations on the ridge south of 36° 17′N (where multibeam bathymetry was not available) were extrapolated from the multibeam data using the seismic reflection data as a guide. Note that measures of Γ in the ridge province may have bias errors due to range-dependent bathymetry and range-dependent α (see Sec. II C 5). Also, the reverberation on the ridge tends to be more sensitive to α in the basin rather than the true local value of α on the ridge (see Sec. II E 7). This suggests that if the bias errors are small, Γ in the ridge province (in particular, close to the province boundary) may be considered as roughly the ratio of scattering strengths on the ridge to that in the basin.

The distribution of Γ for several frequencies (400 Hz, 800 Hz, and 1500 Hz) are shown in Fig. 16 for two pings spaced 20 min and 2.5 km apart, which provide an indication of the stability of Γ over time and space. There is roughly a 90% overlap of the area probed by the two pings. Thus, it is possible to define two provinces over a large area, ~1000 km^2, on the basis of several pings. The boundary between the provinces was obtained by simple thresholding (requiring spatially contiguous levels greater than three standard deviations above the median). Figure 17 shows the two provinces; the precise boundaries were taken from thresholding the 800 Hz data, but the province boundaries are rather insensitive to frequency. Though the boundary between the two provinces has been probed by these few pings, the province extents are presumably larger since the northern, southern, western (for the basin province), and eastern (for the ridge province) boundaries have not been identified. The high variability in the ridge province suggests that it may be divisible into two or more provinces. This division is not explored here.

In the basin province, several discrete scatterers with high levels (greater than three standard deviations) above the background were observed. These can be easily seen in Fig. 9(b); note the three prominent scattered returns at a range of roughly 10 km and beam directions of 75° and 110° and at a range of 3 km and 130°. The geographic position of these scatterers is nearly constant from ping to ping with standard deviations of roughly 400 m obtained by averaging over multiple pings. Other bias errors associated with heading, array location, and range are not known. However, the locations of the scatterers correspond closely to known locations of wrecks (see Table I). More information on discrete scatterers in this area can be found in Ref. 20, where reverberation from a coherent source was processed to produce a scattering map from which fourteen different features are identified and discussed (including these three wrecks).

V. DISCUSSION AND SUMMARY

We have explored the potential for measuring seabed variability using directional reverberation data using simple energy flux models. While this approach is not completely general, (i.e., is only valid for certain classes of sound speed profiles), one of the main contributions of the paper is to show what kinds of variability can and cannot be sensed from reverberation data. Many of the assumptions, sensitivities, errors, and uniqueness issues discussed here may be applicable to other approaches using reverberation data. For

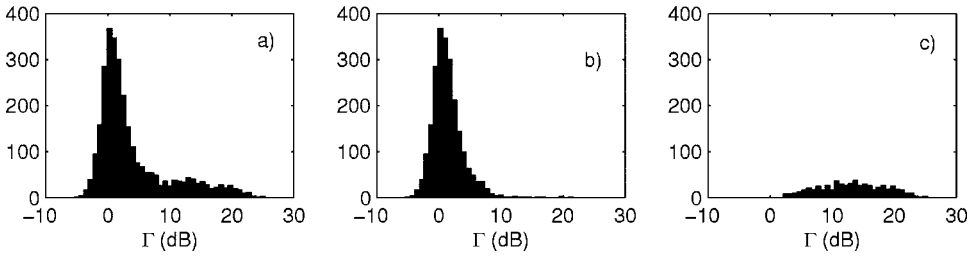


FIG. 14. Histogram of the variation of the scaled reverberation ratio Γ for a single ping at 800 Hz; (a) the entire time series and (b) the area sampled west of the Ragusa Ridge, and (c) Ragusa Ridge. The standard deviation of Γ west of the ridge is 2.3 dB.

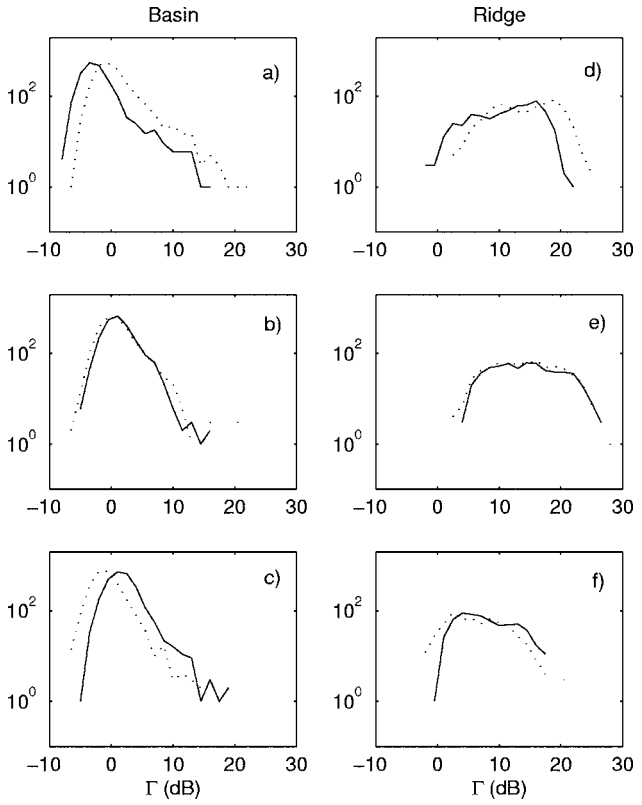


FIG. 16. Comparison of the distribution of Γ as a function of province, Basin (a)–(c) and Ridge (d)–(f) and frequency (a, c) 400 Hz; (b) and (e) 800 Hz; and (c) and (f) 1500 Hz. The solid and dotted lines are the first and fourth pings, separated by about 20 min and 2.5 km.

example, errors were quantified associated with the lack of knowledge of the scattering kernel; this is important because the lack of *a priori* knowledge of the scattering kernel in these types of problems is almost universal. As another example, it was shown that it is not possible to recover an unbiased estimate of the range dependence of the reflection slope α (or equivalently the geoacoustic properties) in the presence of range dependence in the bathymetry. This would be true not only for isovelocity case but also for general types of profiles. Strategies for exploiting the geometry to reduce ambiguity in parameter variability were also developed that may be useful generally.

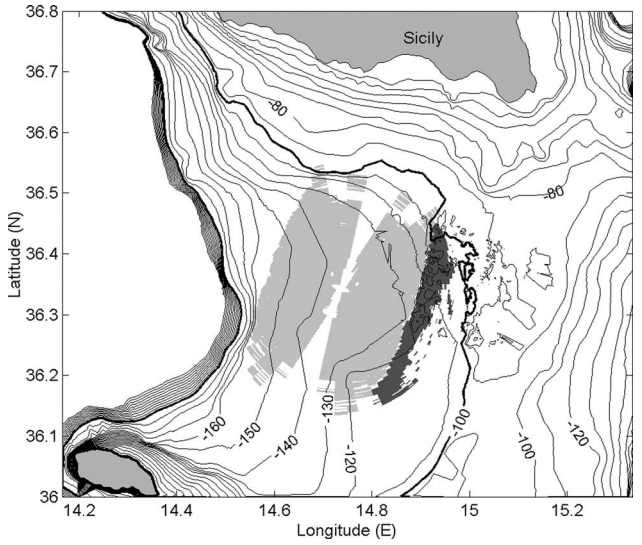


FIG. 17. Two provinces derived from 800 Hz data. Dark gray indicates the ridge province and light gray, the basin province. The results of three pings [spaced 0, 3, and 5 km apart] are plotted and show consistency in the position of the province boundary.

Another important result was the application of the method to measured data. The results showed areas with distinct statistics in the seabed parameters that could be considered as provinces. Two provinces were apparent, and three attributes of the provinces were described including the province boundaries, the statistical distributions of Γ, and the location of discrete scatterers. In the basin province, the distribution was quite narrow, the standard deviation was ~2.5 dB across the entire frequency range (300–1600 Hz). The standard deviation provides some measure of uncertainty of the predictions based on a single (frequency-dependent) value of α and μ for the entire province. In other words, given some fixed value of α and μ, one could expect that measured reverberation would be within ±2.5 dB of the predictions across the province. In the ridge province, the distribution was considerably broader with a standard deviation of 6 dB.

The most obvious application of the provinces would be for a seabed database supporting the same active sonar system that produced the reverberation measurements. This is sometimes referred to as "through-the-sensor" environmental mapping. For active sonar, the statistical distributions provide a measure of the uncertainty of reverberation predictions and perhaps useful metrics for predicting the probability of a false alarm.

In addition, the provinces and the associated statistics also have some utility for passive systems, in some instances, i.e., in cases where the effects of α and μ can be separated. For example, in the basin province, the variability was dominated by variability in μ. (Recall that the speckle must arise from μ and not α.) Thus, passive prediction would have even lower uncertainty. Spatial filtering algorithms could be applied to Γ in the basin province to compute the background variability (i.e., removing the speckle) and so obtain a better estimate of the distribution, but for purposes of this paper it suffices to note simply that the standard deviation for passive systems would be somewhat smaller than the 2.5 dB for active systems.

In addition to usefulness for databases, this provincing approach may be useful for designing a environmental sampling strategy within an area. In other words, analysis of a rapid reverberation survey could potentially identify regions of low and high spatial variability and aid in the design of targeted high-resolution approaches to seabed mapping.

It is remarkable that the seabed is apparently quite homogeneous over an area ~1000 km^2 (i.e., the basin province): homogeneous meaning that the standard deviation of Γ is small, 2.5 dB, and for passive applications would be smaller yet. At first blush, this seems counter to the prevailing notion that shallow-water environments possess extreme variability in the seabed. However, the ostensible homogeneity of Γ in range and azimuth does not mean that the seabed itself is homogeneous. Indeed, seismic records (see Fig. 6), core data, and other geophysical data all show clear hetero-

TABLE I. Location of persistent high scattering events west of the Ragusa ridge and correlation with known wreck positions. In the reverberation data, the median position was taken.

Scatterer locations >3 std from median (Γ)	Known wreck positions	Range offset (m)	Approximate wreck length (m)
36°15.85′ N	36°15.81′ N	160	65
14°48.20′ E	14°48.30′ E		
36°18.52′ N	36°18.28′ N	440	100
14°50.50′ E	14°50.48′ E		
36°18.81′ N	36°18.80′ N	480	Unknown
14°41.46′ E	14°41.14′ E		

geneity over the basin province. What the results here do suggest is that certain kinds of geophysical (and accompanying geoacoustic) variability are sometimes averaged in such a way as to make the seabed appear approximately homogeneous for long-range active systems. The background variability within the province also does not show any indications of range dependence in Γ, i.e., the median levels were essentially independent of range and azimuth. This suggests (see Sec. II C 3) that both μ and α are essentially constant over the basin. The scale of this province was at least 1000 km^2, but the western, southern, and northern edges of the province were not determined so it may be much larger. Whether the size of this province is unusually large in shallow water or rather common is not known.

ACKNOWLEDGMENTS

The author gratefully acknowledges the NATO Undersea Research Centre (under whose auspices the experiments were conducted) and the Office of Naval Research Ocean Acoustics Program (Code OA321) whose support made this analysis possible. The locations of the wrecks were determined in various surveys conducted by Martin Siderius, Mark Prior, and Peter Nielsen.

[1]J. Zhou, D. Guan, E. Shang, and E. Luo, "Long-range reverberation and bottom scattering strength in shallow water," Chinese J. Acoust., **1**, 54–63 (1982).
[2]S. Blanc, J. C. Novarini, and A. L. Nunez, "Bottom reverberation-derive scattering strengths in shallow water off the Argentinian coast," J. Acoust. Soc. Am. **63**, 1342–1346 (1978).
[3]J. R. Preston, D. D. Ellis, and R. C. Gauss, "Geoacoustic parameter extraction using reverberation data from the 2000 Boundary Characterization Experiment on the Malta Plateau," IEEE J. Ocean. Eng. **30**, 2005 (in press).
[4]F. Li, J. Liu, and R. Zhang, "A model/data comparison for shallow-water reverberation," IEEE J. Ocean. Eng. **29**, 1060–1066 (2004).
[5]C. H. Harrison, "Closed-form expressions for ocean reverberation and signal excess with mode stripping and Lambert's law," J. Acoust. Soc. Am. **114**, 2744–2756 (2003).
[6]D. E. Weston, "Intensity-range relations in oceanographic acoustics," J. Sound Vib. **18**, 271–287 (1971).
[7]C. W. Holland, "Shallow water coupled scattering and reflection measurements," IEEE J. Ocean. Eng. **27**, 454–470 (2002).
[8]C. W. Holland, R. Hollett, and L. Troiano, "A measurement technique for bottom scattering in shallow water," J. Acoust. Soc. Am. **108**, 997–1011 (2000).
[9]C. W. Holland, "Rapid geoparameter mapping using local and long-range measurements," (unpublished).
[10]C. H. Harrison, "Closed form bistatic reverberation and target echoes with variable bathymetry and sound speed," IEEE J. Ocean. Eng. **30**, 2005 (in press).
[11]K. D. LePage and C. H. Harrison, "Bistatic reverberation benchmarking exercise: BiStaR versus analytic formulas," J. Acoust. Soc. Am. **113**, 2333–2334 (2003).
[12]K. D. LePage and C. H. Harrison, "Effects of refraction on the prediction of bistatic reverberation in range dependent shallow water waveguides," J. Acoust. Soc. Am. **114**, 2302 (2003).
[13]D. E. Weston, "Propagation in water with uniform sound velocity but variable-depth lossy bottom," J. Sound Vib. **47**, 473–483 (1976).
[14]G. J. Frisk, J.Doutt, and E. Hayes, "Bottom interaction of low-frequency acoustic signals at small grazing angles in the deep ocean," J. Acoust. Soc. Am. **69**, 84–94 (1981).
[15]S. K. Mitchell and K. Focke, "New measurements of compressional wave attenuation in deep ocean sediments," J. Acoust. Soc. Am. **67**, 1582–1589 (1980).
[16]C. W. Holland, "On errors in estimating bottom scattering strength from acoustic reverberation," J. Acoust. Soc. Am. **118**, 2787–2790 (2005).
[17]M. D. Max, A. Kristensen, and E. Michelozzi, "Small-scale Plio-Quaternary sequence stratigraphy and shallow geology of the west-central Malta Plateau," SACLANT Centre Report, SR-209 (1993).
[18]J. Osler and O. Algan, "A high-resolution seismic sequence analysis of the Malta Plateau," NATO Undersea Research Centre, Report SR-311, (1999).
[19]C. W. Holland, R. Gauss, P. Hines, P. Nielsen, D. Ellis, J. Preston, K. D. LePage, C. Harrison, J. Osler, R. Nero, and D. Hutt, "Boundary characterization experiment series overview," IEEE J. Ocean. Eng. **30**, 2005 (in press).
[20]M. K. Prior, "A scatterer map for the Malta Plateau," IEEE J. Ocean. Eng. **30**, 2005 (in press).

Haro Strait geometry (flat bottom)

A. Tolstoy[a)]
ATolstoy Sciences, 1538 Hampton Hill Circle, McLean, Virginia 22101

(Received 17 August 2005; revised 23 November 2005; accepted 29 November 2005)

This paper presents results which relate to the *geometry* of various examined Haro Strait scenarios. In particular, for each *data* scenario the source range, source depth, individual phone ranges, individual phone depths, and an average water depth may be determined based on the time domain signals *alone*. The time differences for each phone for (1) surface reflected minus direct arrivals, and (2) bottom reflected minus direct arrivals are used to estimate potential positions from elementary geometric considerations [similar to Michalopoulou, and Ma (unpublished)]. Nonuniqueness is still a nasty issue and is examined here. However, additional constraints on each scenario to guarantee realism, e.g., the array must be quasivertical with phones 6.25 m apart with no extreme snaking of the array, sources can only be within certain ranges and depths, the top array phone depth and average water depths can only be within certain regimes, etc., eliminate many false solutions. Finally and convincingly, some of the optimal positions are "confirmed" by simulating the signals for the predicted geometries by means of a pulse PE (ramgeo) propagation code (courtesy of M. Collins) and then comparing those simulated signals with the observed data. Once the scenario geometry has been estimated, proper geoacoustic inversion for bottom properties can proceed via the SUB-RIGS method [Tolstoy, IEEE J. Ocean. Eng. **29**, 59–77 (2004(b))]. These subsequent inversions will be discussed in a follow-up paper.
 [DOI: 10.1121/1.2161430]

PACS number(s): 43.30.Pc, 43.60.Jn [EJS] Pages: 1388–1395

I. INTRODUCTION

The Haro Strait experiment south of Vancouver Island conducted in June 1996 was designed to examine acoustic signals received from numerous distributed sources on several (three) essentially vertical arrays in a range-dependent shallow water region. The topography, nominal source, and nominal array positions are shown in Fig. 1 with certain shots emphasized (14,24,29). The emphasized shots will be examined in this paper, and the nominal values are given in Table I. The test data have been studied by a number of researchers (Jaschke and Chapman, 1999; Pignot and Chapman, 2001; Michalopoulou and Ma, 2005) and are generally considered to be extremely difficult to invert as a result of strong local currents as well as complicated region topography. One simplification of the Haro Strait data seems to be the generally constant sound-speed profile (1482.5 m/s), in range and depth, for the entire region at the time of the test. In addition to studies with the test data, efforts have concerned analysis of the data simulated to resemble the Haro Strait data as well (Corré and Chapman, 2001; Tolstoy, 2004a, 2000). The primary intent of all of the analysis has been to estimate the geometric and/or environmental, e.g., bottom, properties of the region. Most recently, efforts to examine data have suggested that scenario geometry is extremely difficult to estimate at single frequencies (Tolstoy, 2005), and the correct geometry is vital to successful geoacoustic inversion. Thus, we will attempt to develop a useful method to determine Haro Strait scenario geometries.

The structure of this paper is as follows: Section II provides a brief presentation of the geometric relations used in the time domain inversions for source range and depth, and for the array phone depths and ranges. Section III discusses the pulse simulations for the inversion values obtained in Sec. II. Section IV presents the pulse simulation results and compares them with the data for three selected paths. Section V summarizes the work and presents conclusions.

II. GEOMETRIC RELATIONS

There is an obvious and simple relationship (also shown in Michalopoulou and Ma, 2005) between the time (T_d) of a direct arrival and a time (T_S) for the first surface reflected arrival for a source in a *constant* sound-speed ocean when modeled by ray acoustics [Skarsoulis and Kalogerakis (2005) have been studying positioning via arrival times assuming a more complicated depth dependent sound-speed profile]. The geometry is shown in Fig. 2(a). In particular, we see that for c_0 as the (assumed) constant sound-speed of the ocean estimated times are given by

$$\hat{T}_d = d/c_0,$$

$$\hat{T}_S = (d_1 + d_2)/c_0,$$

where rge is the range from the source to the phone, zsou is the source depth, and zph is the depth of the phone so that

$$d^2 = \mathrm{rge}^2 + (\mathrm{zsou} - \mathrm{zph})^2,$$

$$(d_1 + d_2)^2 = \mathrm{rge}^2 + (\mathrm{zsou} + \mathrm{zph})^2.$$

Thus, we can easily compute an estimated time difference

[a)]Electronic mail: atolstoy@ieee.org

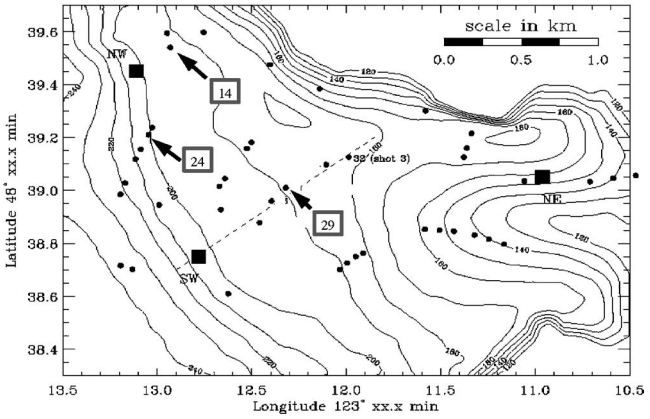


FIG. 1. Nominal values for Haro Strait topography, source, and array positions. Selected shots (14,24,29) will be examined in this paper. The sources are indicated by the black circles while the arrays are indicated by the larger black squares. Figure courtesy of N. Chapman.

TABLE I. Nominal array and source geometric values.

Array	(lat, long) (deg)	zph1 (m)	zph8 (m)
NW	(48.657, 123.217)	30.0	73.75
SW	(48.644, 123.213)	30.0	73.75
Shot No.	**(lat, long) (deg)**	**zsou (m)**	**rge (m)**
14	(48.659, 123.216)	70	215 to NW
24	(48.654, 123.218)	70	380 to NW
29	(48.650, 123.205)	70	887 to SW

$$\hat{\Delta}_S = \hat{T}_S - \hat{T}_d$$

for a given source and phone position. This difference can be compared to the measured data difference ($\Delta_S = T_S - T_d$) to optimize the estimated phone and source positions, i.e., we can minimize $|\hat{\Delta}_S - \Delta_S|$ as a function of zsou, zph, and rge.

Similarly, we can compute the estimated time ($\hat{T}_B$) of a first average bottom reflected arrival for a flat bottom [as seen in Fig. 2(b)], then compute the time difference $\hat{\Delta}_B = \hat{T}_B - \hat{T}_d$ for a variety of bottom depths D, and finally compare the estimated time differences with that observed in the data ($\Delta_B = T_B - T_d$) to find the value of D which minimizes $|\hat{\Delta}_B - \Delta_B|$. We have

$$\hat{T}_B = (d_3 + d_4)/c_0,$$

$$(d_3 + d_4)^2 = \mathrm{rge}^2 + (D - \mathrm{zsou} + D - \mathrm{zph})^2,$$

$$\hat{\Delta}_B = \hat{T}_B - \hat{T}_d.$$

In Fig. 3 we see a typical distribution of time differences for the *surface* reflection situation, i.e., $|\hat{\Delta}_S - \Delta_S|$, for various possible phone depths (zph) for the top array phone depth (zph1) and source depth (zsou) for a nonspecified scenario (rge is fixed). Unrealistic parameter values have been included in Fig. 3 to more clearly show the structure of the optimization solution. The minimum difference (optimal value) here is approximately $2.7 * 10^{-4}$ s which can change as a function of rge (as seen in the formulas). While such a three-dimensional surface is fairly benign (easy enough to find the minimum), and the degradations are not severe as a function of rge, there is clearly an infinite number of excellent positions for even this rge. Minima of 10^{-3} s or less

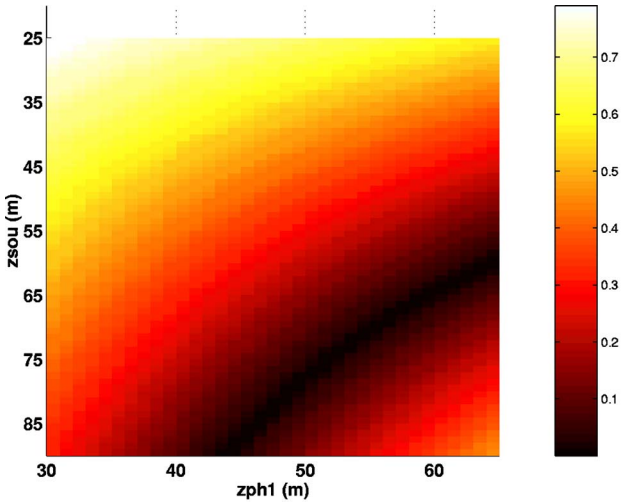


FIG. 3. (Color online) Plot of time differences $|\hat{\Delta}_S - \Delta_S|$ for surface vs direct signal as a function of zsou and zph for a single phone. Scenario is unspecified but is at fixed range and involves Haro Strait data.

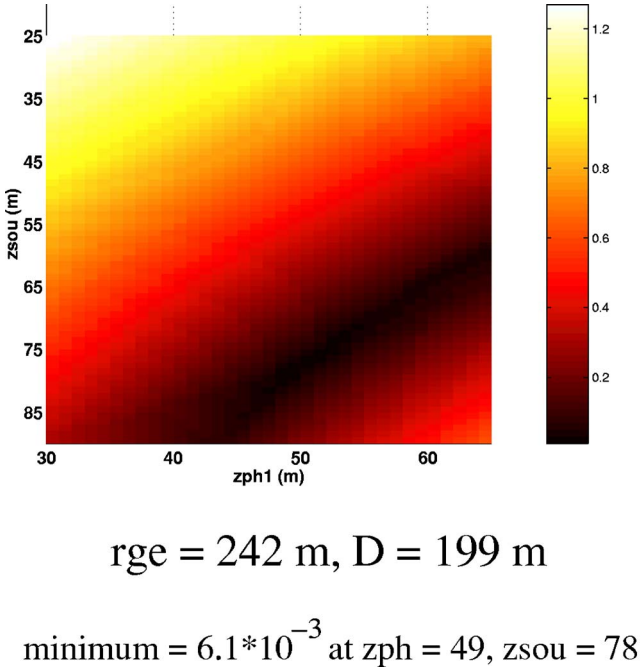


FIG. 4. (Color online) Plot of time differences $|\hat{\Delta}_S - \Delta_S| + |\hat{\Delta}_B - \Delta_B|$ for first surface vs direct signal *plus* first bottom vs direct signal as a function of zsou and zph for a single phone. Scenario is unspecified but is at fixed range and fixed water depth and involves Haro Strait data.

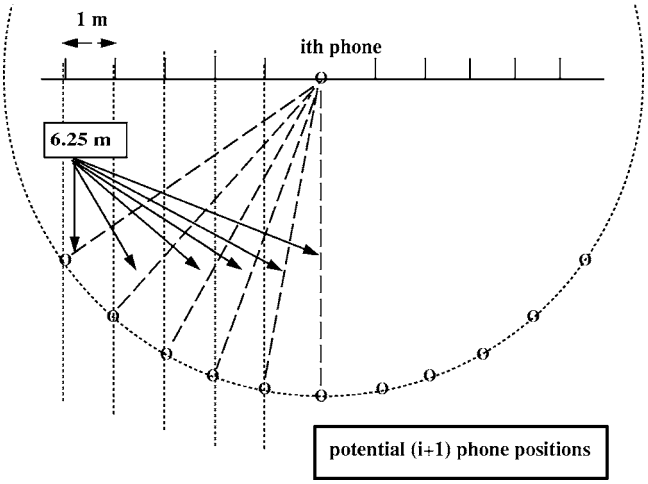


FIG. 5. Plot showing the potential phone positions for the $(i+1)$th phone relative to the ith phone. The $(i+1)$th phone must be deeper than the ith phone but can be either toward the source (5 possible discrete positions including vertical), away from the source (5 other possible discrete positions including vertical), or vertically below the higher phone. The phone must be 6.25 m distant from the above phone.

result in essentially perfect agreement for Haro Strait data where the time sampling is 0.000 57 s.

Let us extend the above-mentioned computation to include consideration of *bottom* effects, i.e., to include $|\hat{\Delta}_B - \Delta_B|$. Results are seen in Fig. 4 (parameters corresponding to Fig. 3), but where water depth is fixed $\hat{D}=199$ m. Here the minimum value (sum) has increased to $6.1*10^{-3}$. The choices of very good fits to the data are still enormous.

If we extend this process to include an *array* of phones we find that some realizations can be very unrealistic. If we place restrictions on the array phone positions relative to each other, we can eliminate many false configurations. In Michalopoulou and Ma the restrictions are done via regularization. Here, they are done by limiting the possible configurations of each phone to a discrete set of possible displacements from the above phone (see Fig. 5). We consider arrays

TABLE II. Optimal source depth (zsou), phone locations (zph, rge), and water depth D given Haro Strait data for a selected path (nw014).

zsou=77.0 m in [60,80] Phone depth (m)	D=199.0 m in [185,210] Phone range (m)
zph_1=49.75 m in [20,50]	rge_1=242.00 m in [150,300]
zph_2=53.50 m	rge_2=247.00 m
zph_3=59.23 m	rge_3=249.50 m
zph_4=64.96 m	rge_4=252.00 m
zph_5=71.08 m	rge_5=253.25 m
zph_6=77.20 m	rge_6=254.50 m
zph_7=83.33 m	rge_7=255.75 m
zph_8=89.45 m	rge_8=257.00 m

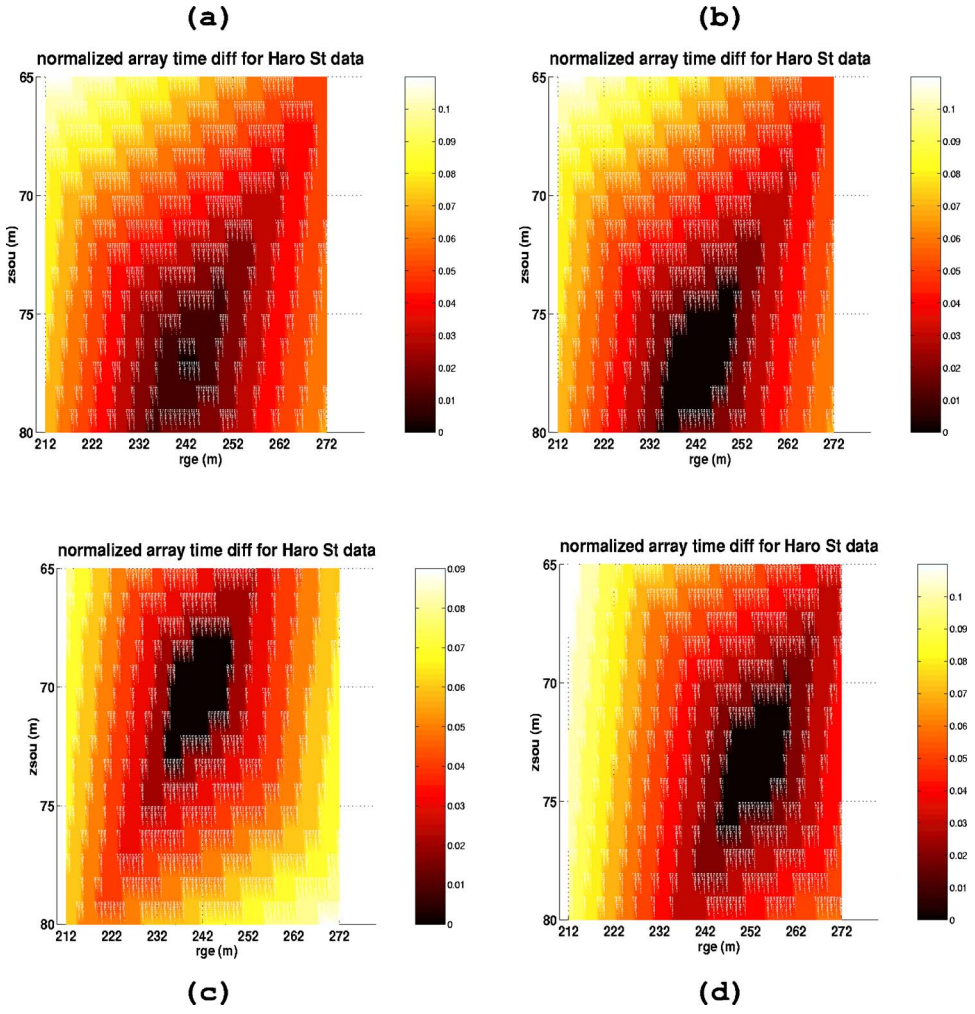


FIG. 6. (Color online) The total time difference agreement with *array* data (for both the first bottom and surface reflections) as a function of zsou and rge. The array depths and deformation are as in Table II (nw014), and D =199 m. The minimum value is at rge=242 m, zsou=77 m seen clearly in (a). In (b) we see the same results but with the "excellent" values indicated by the black coloration. In (c) we see more excellent values but for a simple *vertical* array with top phone at 50 m. Finally, in (d) we see more excellent values for the vertical array but with D=195 m.

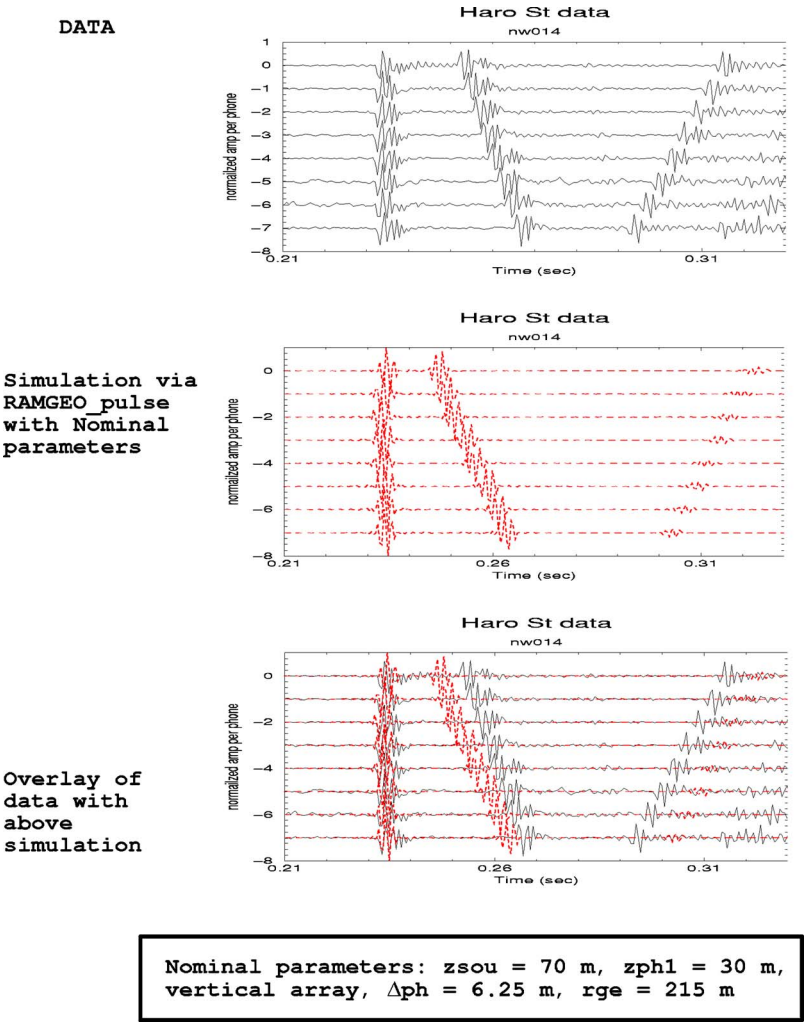


FIG. 7. (Color online) Data and simulated pulse fields at the NW array for the path from shot 14. (a) The data as a function of time and phone (the deeper phones are the higher negative numbers); (b) the field simulated via the pulse ramgeo for the nominal parameter values; (c) overlay of the data and simulated fields. The simulated data have been shifted to line up with the direct arrivals.

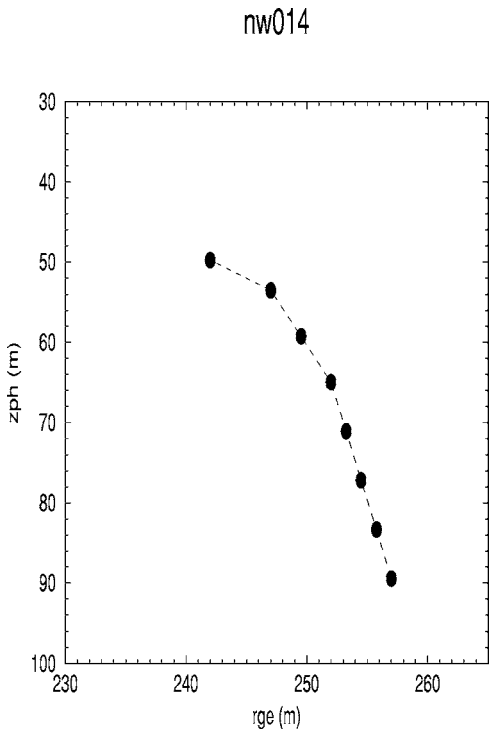


FIG. 8. Optimized nw014 array shape for a "tilted" array. Phone depths and ranges from the source are as indicated. The array length is 43.75 m.

which are either tilted toward the source (five of the indicated potential positions), tilted away from the source (the other five of the indicated positions), or vertical. Subsequently, we can compute a family of phone ranges and depths for various arrays which again result in excellent agreement between time differences for the simulated and the measured time domain fields. In Table II we see one such result (a best array, a best water depth, and a best source fit to the data) for a particular Haro Strait path (nw014).

Finally, with regard to the nonuniqueness of the results, we illustrate the choices for source range (to the first array phone) and source depth given the best values of Table II

TABLE III. Optimal source depth (zsou), phone locations (zph, rge), and water depth D given Haro Strait data for a selected path (nw024).

zsou=69.0 m in [60,80] phone depth (m)	D=185.0 m in [185,210] phone range (m)
zph_1=36.00 m in [30,60]	rge_1=412.00 m in [350,450]
zph_2=41.00 m	rge_2=415.75 m
zph_3=46.00 m	rge_3=419.50 m
zph_4=51.00 m	rge_4=423.25 m
zph_5=57.25 m	rge_5=423.235 m
zph_6=61.00 m	rge_6=428.25 m
zph_7=66.73 m	rge_7=430.75 m
zph_8=72.98 m	rge_8=430.75 m

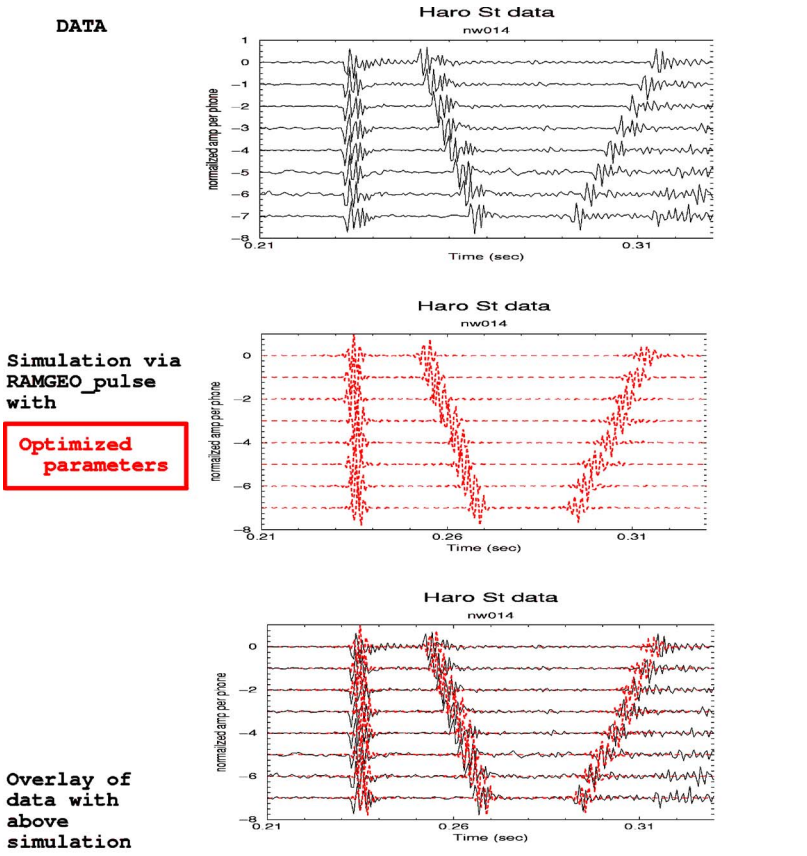


FIG. 9. (Color online) Data and simulated pulse fields at the NW array for the path from shot 14. (a) The data as a function of time and phone [as in Fig. 6(a)]; (b) the field simulated via the pulse ramgeo for the optimal parameter values of Table II; (c) overlay of the data and simulated fields. The simulated data have been shifted to line up with the direct arrivals.

Optimized parameters: zsou = 77 m, zph1 = 50 m, deformed/tilted array, Δph = 6.25 m, rge = 242 m, D = 199.0 m.

NOTE: cbot = 2400 m/s

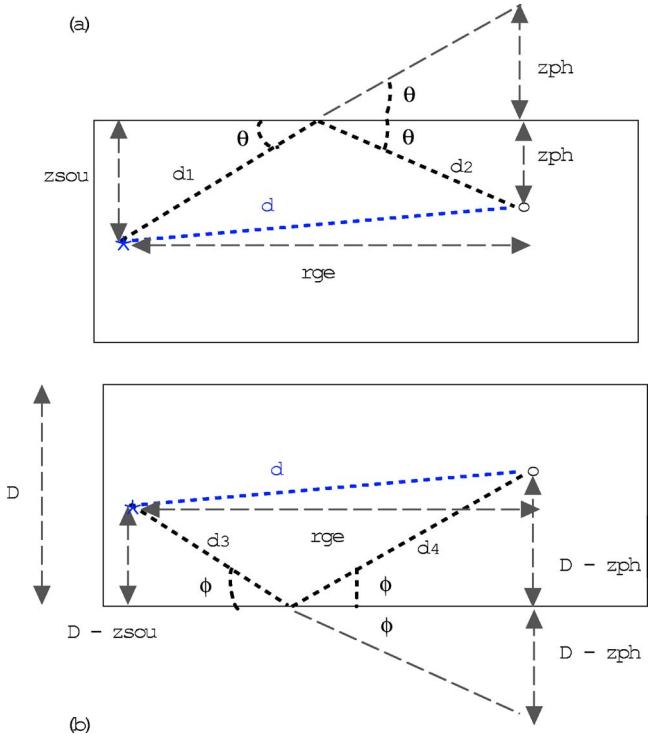


FIG. 2. (Color online) Geometric relationships between (a) the first surface reflected arrival distance (d_1+d_2) and the direct arrival distance d; (b) the first bottom reflected arrival distance (d_3+d_4) and the direct arrival distance d.

(array shape and water depth) and given a simpler vertical array. In Fig. 6(a) we see the minimal value clearly at zsou =77 m, rge=242 m. In subsequent parts we have indicated all the "excellent" values which would agree with the data with nearly perfect accuracy (time differences less than a total of 0.016 s for 8 phones, surface plus bottom reflections) by the black coloration. In Fig. 6(b) we see that there are numerous options for such excellent values given our array and D=199 m. The minimum (best agreement to data) remains at rge=242 m, zsou=77 m. However, many other values will do quite well. In Fig. 6(c) we see that a vertical array with the top phone at 50 m and D=199 m results in other excellent values. In Fig. 6(d) we have the same vertical array but D=195 m. Thus, there are many excellent values of the geometric parameters.

III. PULSE SIMULATIONS

Next, we proceed to examine the data versus the pulse simulated time fields using optimized values computed as discussed earlier for a selection of Haro Strait paths. The propagation model which we are using is the pulse ramgeo (Collins, 1994) PE model given to us by Collins. The source spectrum which we are assuming does not exactly match the source spectrum of the data [a typical wavelet is seen in Pignot and Chapman (2001), while the original spectrum and

a slightly different wavelet is seen in Jaschke (1997)].[1] The assumed spectrum is a very good approximation (assuming a Gaussian source centered on 600 Hz with a standard deviation of 200 Hz[2]). The results consider frequencies from 200 to 800 Hz at 1-Hz intervals (a total of 600 frequencies). For all the results to follow, the *nominal* arrays were considered to be vertical with phone distances of 6.25 m. We shall consider only the top eight (8) phones for processing (maximum array length of 43.75 m).

A. Test case nw014

Examining Fig. 1 we see the path from source 14 to the NW array (nw014). This path has nominal values: zsou =70 m, rge=215 m, zph1=30 m, with the bottom depth at the source at 190 m, the bottom depth at the array of 205 m. In Fig. 7 we see the data measured at the phones (a and c) as well as at the simulated field (b and c) assuming the nominal values and a bottom sound-speed of 1600 m/s. We note that the agreement is not very good for the nominal values. Additionally, the CPU time for each PE run is approximately 2 h. Not only are the differences between the direct and the surface reflected arrivals poor, but the simulated bottom arrivals are so faint as to be nearly nonexistent suggesting that optimization efforts should allow for a faster bottom or subbottom with a stronger reflection coefficient.

After optimization to find the minimum $|\hat{\Delta}_S - \Delta_S| + |\hat{\Delta}_B - \Delta_B|$ over all "reasonable" phone deformations, all likely first phone depths [20,50], all likely source depths [65,80], all likely ranges to the first phone range in [150,300], and all likely "average" bottom depths D in [185,210], we obtain the values of Table II. The optimized array shape is shown in Fig. 8 where we have considered only arrays with no snaking, i.e., only "tilted" arrays. It should be noted that a deeper, slightly different shaped array (top phone at 54.0 m) at a farther range (257 m) would result in even better fits to the data. However, such a deep array depth has been rejected as unrealistic.

Using the optimal parameters of Table II with a bottom sound-speed of 2400 m/s (not realistic but guaranteeing a reflection coefficient of 1.00[3]), we obtain the simulated fields of Fig. 9 where we see very good agreement with the data for this path. The agreement is not perfect, however. This may be the result of a nonflat bottom for the data. Future efforts to determine the geoacoustic and topographic properties of the bottom via matched field tomographic inversion should improve this as well as improve later arrival agreements (none of which are shown here).

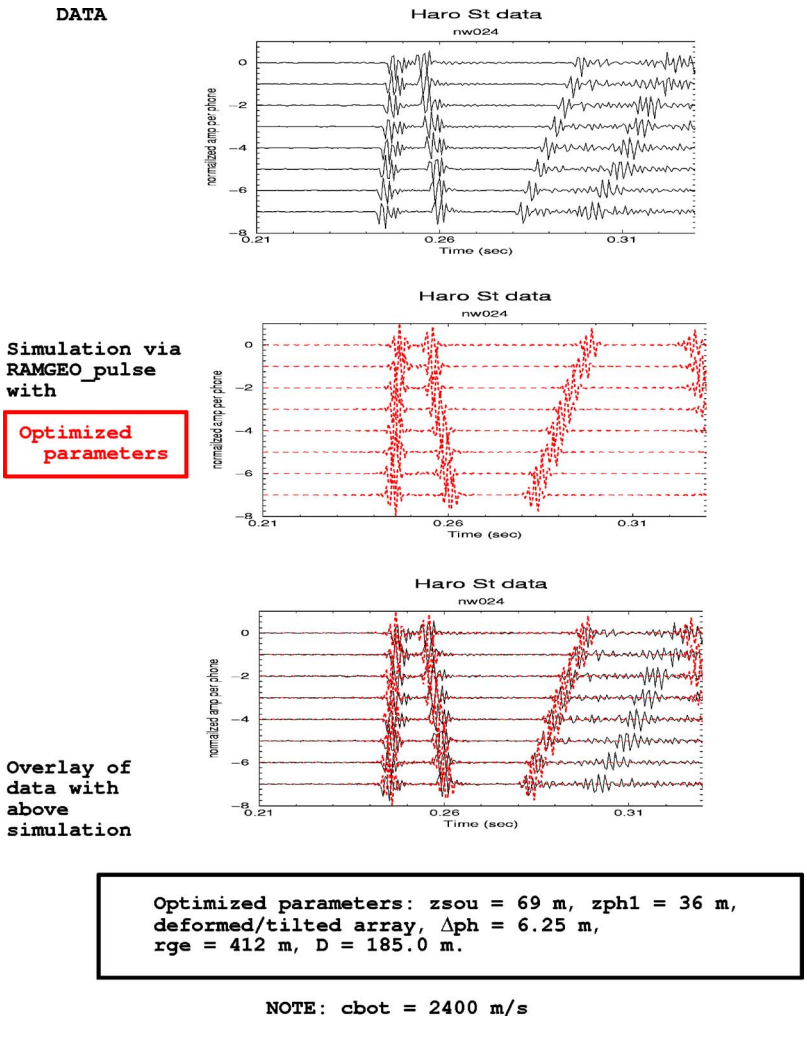


FIG. 10. (Color online) Data and simulated pulse fields at the NW array for the path from shot 24. (a) The data as a function of time and phone; (b) the field simulated via the pulse ramgeo for the optimal parameter values of Table III; (c) overlay of the data and simulated fields. The simulated data have been shifted to line up with the direct arrivals.

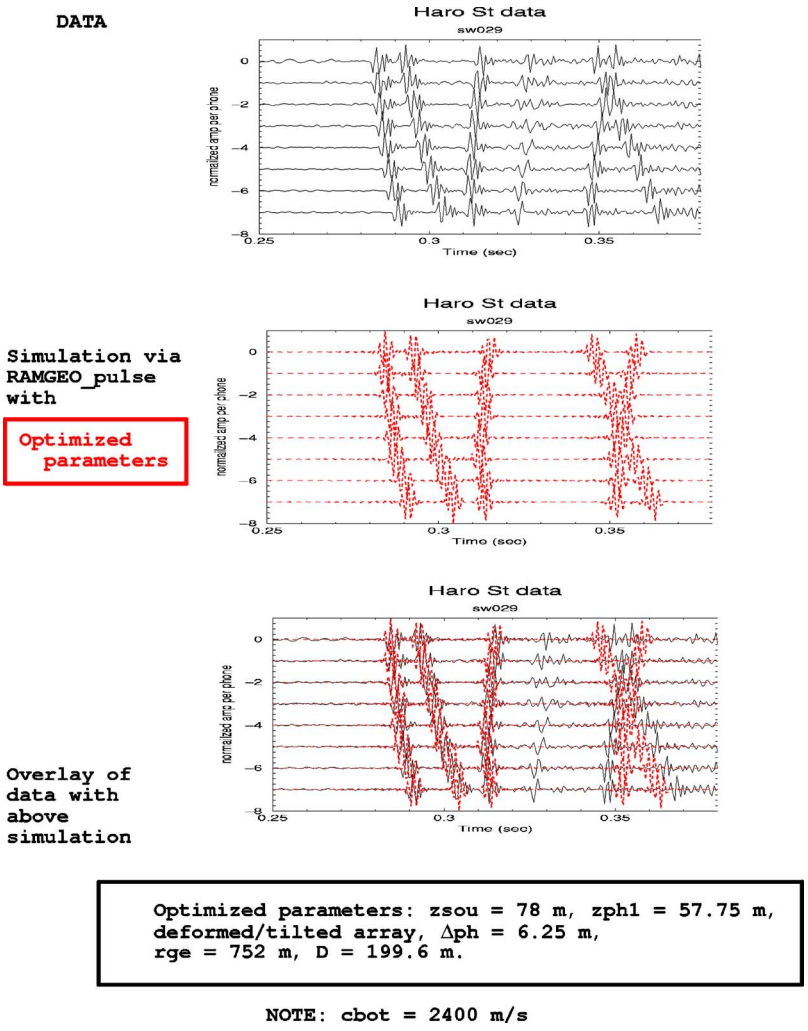


FIG. 11. (Color online) Data and simulated pulse fields at the SW array for the path from shot 29. (a) The data as a function of time and phone; (b) the field simulated via the pulse ramgeo for the optimal parameter values of Table IV; (c) overlay of the data and simulated fields. The simulated data have been shifted to line up with the direct arrivals.

B. Test case nw024

For the next test case nw024, we arrive at the optimal parameters shown in Table III. The agreement with the data is excellent, and the data versus optimal simulated fields shown in Fig. 10. The CPU time for each PE run is approximately 4 h (nominal values are shown in Table I). Simulated fields for the nominal values are not shown. We note that the NW array configuration is quite different, e.g., phone depths, from that of the previous scenario (nw014). We recall that the currents can be quite different at this time versus the last time and will result in very different effects on the array. Still the very different depths of this prediction versus those of case nw014 using the same array are troublesome. Either the results are real, indicating extremely strong currents, or some assumption is inappropriate, e.g., the bottom is not flat and has a major effect, the sound-speed is not constant, or the estimated time intervals used in the computations are incorrect. This discrepancy has been noted by one of the reviewers, is important, and will be addressed in future work.

C. Test case sw029

For the next test case sw029, we arrive at the optimal parameters shown in Table IV. The agreement with the data is again excellent, and the data versus optimal simulated fields are shown in Fig. 11. The CPU time for each PE run is approximately 5 h (nominal values are shown in Table I). Simulated fields for the nominal values are not shown.

IV. CONCLUSIONS

We have demonstrated in this paper that geometric considerations can suggest optimal source and array configurations to match the observed data over a broad range of frequencies (200–800 Hz). We have illustrated these matches via a pulse ramgeo (PE) model for parameter values found by simple straight line geometry. However, there are *many* possible arrangements which can result in excellent agree-

TABLE IV. Optimal source depth (zsou), phone locations (zph, rge), and water depth D given Haro Strait data for a selected path (sw029).

zsou=78.0 m in [60,80] phone depth (m)	D=199.6 m in [185,210] phone range (m)
zph_1=57.75 m in [30,60]	rge_1=752.00 m in [700,900]
zph_2=64.00 m	rge_2=752.00 m
zph_3=70.25 m	rge_3=752.00 m
zph_4=74.00 m	rge_4=757.00 m
zph_5=80.12 m	rge_5=758.25 m
zph_6=86.37 m	rge_6=758.25 m
zph_7=90.12 m	rge_7=763.25 m
zph_8=95.85 m	rge_8=765.75 m

ment with the data. That is, the nonuniqueness issue is still problematic. This is suggested by examining figures such as Figs. 3–5, and by recalling that other scenarios could be considered for the test cases—particularly if restrictions are changed. Thus, even broadband processing may not be sufficient to completely resolve ambiguity issues, particularly as considered here. However, geometric considerations can significantly reduce the search spaces for such parameters as zsou, rge, zph1, and *D*.

Additionally, Tables II and III suggest very different array depths for different sources (times) but for the same array. These very different depths may be realistic since the region is very dynamic. However, attempts to confine an array to smaller variations in depth will be done in later work in an effort to guarantee more consistency between shots.

We note that matches here to the data depend upon estimates of the arrival times of the direct, surface-reflected, and bottom reflected fields. These are not always simple to do and can introduce discrepancies. Additionally, we assume a flat bottom (average) for our optimizations. This is not realistic, can change the results, and may account for more of the observed differences.

Finally, nonuniqueness may be reduced by consideration of a nonflat bottom, e.g., bottom slope. It may be improved by using additional, higher order reflections in the data such as surface-bottom-reflected and bottom-surface-reflected arrivals. While subbottom layers may be responsible for some of the observed arrivals, it may be possible to tease out the simpler reflections. Nonuniqueness may also be reduced by using more of the array phones than done here. These considerations will be addressed in a future paper.

ACKNOWLEDGMENTS

The author would like to thank ONR for support through Grant No. N00014-05-M-0023 and for direction on the Haro Strait data analysis; R. Chapman for his continued time and efforts to supply and discuss the data from this unique but difficult test data; M. Collins for his excellent pulse PE code and numerous communications concerning its use; and E. Michalopoulou for her very helpful discussions and thoughts. Finally, the reviewers have been very helpful, showing much insight with constructive suggestions which were much appreciated.

[1]The data source is a lightbulb implosion where at 70 m depth the center frequency is 600 Hz with a 3dB bandwidth of 300 *rmHz*.

[2]Frequencies higher than 800 Hz and lower than 200 Hz are neglected since they will not be used in geoacoustic inversions to be done in later work.

[3]More realistic values will be likely after geoacoustic inversion is performed in later work.

Collins, M. D. (**1994**). "Generalization of the split-step Pade solution," J. Acoust. Soc. Am. **96**, 382–385.

Corré, V., and Chapman, N. R. (**2001**). "Vertical-slice matched field tomography," Acoustica **87**, 637–646.

Jaschke, L. (**1997**). "Geophysical inversion by the freeze bath method with an application to geoacoustic ocean bottom parameter estimation," Ph.D. thesis, U. Victoria.

Jaschke, L., and Chapman, N. R. (**1999**). "Matched field inversion of broadband data using the freeze bath method," J. Acoust. Soc. Am. **106**, 1838–1851.

Michalopoulou, Z.-H., and Ma, X. (2005). "Source localization in the Haro Strait using arrival time estimation and linearization," J. Acoust. Soc. Am. (submitted).

Pignot, P., and Chapman, N. R. (**2001**). "Tomographic inversion in a range-dependent shallow water environment," J. Acoust. Soc. Am. **110**, 1338–1348.

Skarsoulis, E. K., and Kalogerakis, M. A. (2005). "Ray-theoretic localization of an impulsive source in a stratified ocean using two hydrophones," J. Acoust. Soc. Am. (submitted).

Tolstoy, A. (**2005**). "Geoacoustic inversion for single paths of Haro Strait data," J. Acoust. Soc. Am. **117**, 2443.

Tolstoy, A. (**2004**a). "Simulated tomographic geoacoustic inversion," in *Theoretical and Computational Acoustics 2003*, ed., pp. 403–413.

Tolstoy, A. (**2004**b). "Matched field processing (MFP)-based inversion method (SUB-RIGS) for range-dependent scenarios," IEEE J. Ocean. Eng. **29**, 59–77.

Tolstoy, A. (**2000**). "Tomographic inversion for geoacoustic parameters in shallow water," J. Comput. Acoust. **8**, 285–293.

Classification of shallow-water acoustic signals via alpha-Stable modeling of the one-dimensional wavelet coefficients

Michael I. Taroudakis[a)]
Department of Mathematics, University of Crete, Institute of Applied and Computational Mathematics-FO.R.T.H., IACM-FORTH, P.O. Box 1527, 711 10 Heraklion, Crete, Greece

George Tzagkarakis[b)] and Panagiotis Tsakalides[c)]
Department of Computer Science, University of Crete, Institute of Computer Science-FO.R.T.H. ICS-FORTH, P.O. Box 1385, 711 10 Heraklion, Crete, Greece

(Received 19 August 2005; revised 13 December 2005; accepted 13 December 2005)

A novel statistical scheme is presented for the classification of shallow water acoustic signals according to the environmental parameters of the medium through which they have propagated. An efficient way to classify these signals is important for inverse procedures in underwater acoustics aiming at the recovery of the geoacoustic parameters of an oceanic environment, using measurements of the acoustic field due to an acoustic source. An important issue in this procedure is the determination of an efficient "observable" of the acoustic signal (feature extraction), which characterizes the signal in connection with the recoverable parameters. The proposed method is based on a transformation of the acoustic signals via a one-dimensional (1D) wavelet decomposition and then by fitting the distribution of the subband coefficients using an appropriate function. We observe that statistical distributions with heavy algebraic tails, such as the alpha-Stable family, are often very accurate in capturing the non-Gaussian behavior of the subband coefficients. As a result, the feature extraction step consists of estimating the parameters of the alpha-Stable model, while the similarity between two distinct signals is measured by employing the Kullback–Leibler Divergence between their corresponding alpha-Stable distributions. The performance of the proposed classification method is studied using simulated acoustic signals generated in a shallow water environment. [DOI: 10.1121/1.2165003]

PACS number(s): 43.30.Pc, 43.60.Pt, 43.60.Lq [AIT] Pages: 1396–1405

I. INTRODUCTION

Among the most interesting inverse problems in underwater acoustics is that of determining the sea-bed or water column parameters from acoustic measurements obtained in the water column. A great amount of literature is devoted to different issues of inversion procedures in underwater acoustics (e.g., Refs. 1–21). Matched-field,[1–7] and modal phase[8–10] inversions are based on measurements made at a number of hydrophones greater than one, whereas ray inversions typical in ocean acoustic tomography[11–13] and modal travel time inversions[14–16] are based on measurements made at a single hydrophone. With the exception of the matched-field techniques, in all other cases a suitable "observable" should be clearly identified. For instance, identification of the modal character of a signal may be based on mode filtering (when multiple receivers are available),[17] on time domain techniques[18] or on techniques utilizing characteristics of the signal in the time-frequency domain.[19–21] The observable is used as the link between the recoverable parameters and the measurements.

All the above mentioned techniques have been the subject of on-going research for the improvement of their performance, in connection with linear or nonlinear inversion procedures, which are necessary for the estimation of a solution to the inverse problem that would be as close to reality as possible. The problem is that the techniques are not always applicable in practical applications (for instance rays or modes may not be identified in a specific experiment), or the sensitivity of the observable to slight changes of the environmental parameters is not always enough for an inversion technique to be applied with confidence and therefore, alternative techniques for signal classification based on a different feature are always in the front of research.

It should be noted that most of the inversion procedures and the associated observable identification (feature extraction) are based on deterministic approaches. The present work concerns signal classification in the case of a single reception, when the available measurement is typically a signal in the time domain. Here, an alternative technique is proposed, in which the tasks of feature extraction (FE) and similarity measurement (SM) are considered in a joint statistical framework. In particular, features that precisely and uniquely describe the internal characteristics of the underwater acoustic signal, are studied taking into account the fact that these features should be sensitive to slight variations of the most important parameters of the environment under consideration. In the proposed approach, the FE step becomes a Maximum Likelihood (ML) estimator of the model parameters fitting the given transformed acoustic signal,

[a)]Electronic mail: taroud@iacm.forth.gr
[b)]Electronic mail: gtzag@ics.forth.gr
[c)]Electronic mail: tsakalid@ics.forth.gr

while the SM step employs the Kullback–Leibler divergence (KLD),[22] which is a statistical measure of similarity between probability density functions having different model parameters. Using this statistical approach each signal is modeled by the marginal densities of the transform coefficients. The above joint statistical framework can then be used in connection with an inversion scheme for the characterization of the ocean environment.

The development of classification schemes in a transform domain is based on the observation that often a linear, invertible transform restructures the signal, resulting in a set of transform coefficients whose structure is simpler to model. For bursty shallow-water acoustic signals, the 1D wavelet transform seems to be a powerful modeling tool, providing a natural arrangement of the wavelet coefficients into multiple scales representing the frequency content of the signal in consecutive bands.[23] Besides, it has been pointed out that the wavelet transforms of bursty signals tend to be sparse, resulting in a large number of coefficients with small magnitude and a small number of large magnitude coefficients.[24] This property gives rise to peaky and heavy-tailed *non-Gaussian* marginal distributions of the wavelet subband coefficients.[24]

The *symmetric alpha-stable* ($S\alpha S$) distributions,[25,26] have proven to be efficient in describing the content of many texture images,[27] which can be considered as of similar type with that of an acoustic signal due to its impulsive character. In the present work, it is demonstrated that this family of distributions is also appropriate for the statistical characterization of a shallow-water acoustic signal. The similarity measurement between two acoustic signals is performed through an appropriate version of the KLD between the corresponding $S\alpha S$ distributions.

The paper is organized as follows: In Sec. II, the mathematical modeling of the source-channel-receiver system is described. In Sec. III, the probabilistic setting for the problem of classifying underwater acoustic signals is briefly reviewed. In Sec. IV, the choice of the univariate symmetric alpha-stable ($S\alpha S$) model for the modeling of the marginal distributions of the wavelet subband coefficients is justified, while the feature extraction as well as the construction of an appropriate similarity measure is described. In Sec. V, the proposed scheme is applied to a database of simulated acoustic signals generated in a shallow-water environment to evaluate the classification performance.

II. MODELING THE SOURCE-CHANNEL-RECEIVER SYSTEM

The acoustic signals to be considered in the present study are simulated broadband tomographic signals modeled using acoustic wave theory applied in shallow water environments. An axially symmetric range-independent environment will be considered.

The acoustic pressure p at a specific location $\mathbf{x}$ for a sound source in location $\mathbf{x}_0$ of frequency ω is given by formula:

$$p(\mathbf{x};\mathbf{x}_0;\omega) = H_{SR}(\mathbf{x};\mathbf{x}_0;\omega)S(\omega), \tag{1}$$

where $H_{SR}(\omega)$ is the system transfer function determined by the Helmholtz equation written in a cylindrical coordinate system (appropriate for the environment under consideration) for a source placed at $r=0$, $z=z_0$;

$$\nabla^2 H_{SR}(r,z;z_0;\omega) + k^2(z)H_{SR}(r,z;z_0;\omega) = -\frac{1}{2\pi r}\delta(r)\delta(z-z_0), \tag{2}$$

supplemented by the appropriate boundary and radiation conditions. $k(z)=\omega/c(z)$ is the wave number, where ω is the circular frequency, and $c(z)$ the sound speed at depth z. $S(\omega)$ is the source excitation function, which is the present study will be considered a Gaussian pulse representing standard types of tomographic signals.

The pressure field in the time domain is given by the inverse Fourier transform:

$$p(r,z;z_0;t) = \mathcal{F}^{-1}[p(r,z;z_0;\omega);\omega\to t] = \frac{1}{2\pi}\int_{-\infty}^{\infty} H_{SR}(r,z;z_0;\omega)S(\omega)e^{i\omega t}d\omega. \tag{3}$$

As in the present study, the analysis is free from any restrictions imposed by the observable, the system transfer function can be calculated using any appropriate propagation model. However, the fact that the environment is considered range independent, a normal mode code is the most accurate for the calculation of the pressure field in the frequency domain and it will be used for the simulation of the acoustic signal.

III. STATISTICAL SIGNAL CLASSIFICATION

In the analysis to follow, signal classification will be associated with an inversion scheme not explicitly defined here, based on minimization of an appropriately defined difference between a measured signal and a signal from a database or from a set of signals simulated from a search space on the environmental parameters to be recovered. The difference has to be determined using a signal feature (or observable). The term "query" used below is in fact the signal to be classified.

Let $\mathcal{F}$ denote the feature space and $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_N|\mathbf{x}_i \in \mathcal{F}, i=1,\ldots,N\}$ be a set of N independent feature vectors associated to a query. Also, let $\mathcal{S}=\{1,\ldots,K\}$ be the set of class indicators associated with the classes in the database, where each class corresponds to a specific sea environment. Besides, in our setting we consider that each class contains one signal generated with the corresponding environmental parameters. Thus, in the rest of the paper a "class" is equivalent to an environment associated to the corresponding signal in the database. Denote the probability density function (PDF) of the query feature vector space by $p_q(\mathbf{x})$ and the PDF of class $i\in\mathcal{S}$ by $p_i(\mathbf{x})$. The design of a classification scheme in a probabilistic framework, consists of finding an appropriate map $g:\mathcal{F}\mapsto\mathcal{S}$. These maps constitute the set of similarity functions.

The goal of a probabilistic classification system is the

minimization of the probability of classification error, that is, the probability $P(g(\mathbf{X}) \neq s)$. Hence, if we provide the system with a set of feature vectors $\mathbf{X}$ drawn from class $s \in \mathcal{S}$, we want to minimize the probability that the system will classify the query in a class $g(\mathbf{X})$ different from s. It can be shown[28] that the optimal similarity function, that is, the one minimizing $P(g(\mathbf{X}) \neq s)$, is the Bayes or maximum a-posteriory (MAP) classifier

$$g_{\text{opt}}(\mathbf{X}) = arg \max_i P(s=i|\mathbf{X}) = arg \max_i p(\mathbf{X}|s=i)P(s=i), \quad (4)$$

where $p(\mathbf{X}|s=i)$ is the likelihood for the ith class and $P(s=i)$ its prior probability. Under the assumption that all classes are a priori equally likely, the MAP classifier reduces to the ML classifier:

$$g_{\text{opt}}(\mathbf{X}) = arg \max_i p(\mathbf{X}|s=i) \overset{i.i.d.}{=} arg \max_i \frac{1}{N}\sum_{j=1}^{N} \log p(\mathbf{x}_j|s=i). \quad (5)$$

When the number N of feature vectors is large, application of the weak law of large numbers[29] to Eq. (5) results in the following equation:

$$g_{\text{opt}}(\mathbf{X}) = arg \min_i \underbrace{\int p_q(\mathbf{x}) \log \frac{p_q(\mathbf{x})}{p_i(\mathbf{x})} d\mathbf{x}}_{D(p_q\|p_i)}, \quad (6)$$

where $D(p_q\|p_i)$ denotes the KLD or *relative entropy* between the two densities, $p_q(\cdot)$ and $p_i(\cdot)$.

The problem of classifying a given query signal can be formulated as a *hypothesis problem*. The query signal S_q is represented by a feature data set, $\mathbf{X}=\{x_1,\ldots,x_N\}$, obtained after a transformation step, and each class in the database, $S_i(i=1,\ldots,C)$, is assigned with a hypothesis H_i. Therefore, the problem of classifying S_q in one of the database classes, and thus recovering the environmental parameters corresponding to this class, consists of selecting the class in the database that is closer in terms of best hypothesis to the data $\mathbf{X}$ of the given query signal S_q.

Under the assumption that all hypotheses are a priori equally likely, which is equivalent to assume that all values from a predefined search space are equally possible, the optimum rule resulting in the minimum probability of classification error, is to select the hypothesis with the highest likelihood among the C. Thus, the selected class corresponds to the hypothesis, H_{i_1} for which

$$p(\mathbf{X}|H_{i_1}) \geqslant \cdots \geqslant p(\mathbf{X}|H_i), \quad i \neq i_1.$$

A computationally efficient implementation of this setting is to adopt a *parametric* approach. Then, each conditional PDF, $p(\mathbf{X}|H_i)$ is modeled by a member of a family of PDFs, denoted by $p(\mathbf{X};\boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}_i$ is a set of model parameters to be specified. In this framework, the extracted signature for the signal S_i is the *set of estimated model parameters* $\hat{\boldsymbol{\theta}}_i$, computed in the FE step. Then, implementation of Eq. (6) gives the optimal rule for classifying the given query signal S_q to the closest class:

(1) Compute the KLDs between the query density $p(\mathbf{X};\boldsymbol{\theta}_q)$ and the density $p(\mathbf{X};\boldsymbol{\theta}_i)$ associated with class S_i in the database, $\forall i=1,\ldots,C$:

$$D(p(\mathbf{X};\boldsymbol{\theta}_q)\|p(\mathbf{X};\boldsymbol{\theta}_i)) = \int p(x;\boldsymbol{\theta}_q)\log \frac{p(x;\boldsymbol{\theta}_q)}{p(x;\boldsymbol{\theta}_i)} dx. \quad (7)$$

(2) Classify S_q in the class corresponding to the smallest value of the KLD.

The KLD in Eq. (7) can be computed using consistent estimators $\hat{\boldsymbol{\theta}}_q$ and $\hat{\boldsymbol{\theta}}_i$, for the model parameters. The ML estimator is a consistent estimator[22] and for the query signal it gives:

$$\hat{\boldsymbol{\theta}}_q = arg \max_{\boldsymbol{\theta}} \log p(\mathbf{X};\boldsymbol{\theta}). \quad (8)$$

A *chain rule*[29] can also be applied, in order to combine the KLDs from multiple data sets. This rule states that the KLD between two joint PDFs, $p(\mathbf{X},\mathbf{Y})$ and $q(\mathbf{X},\mathbf{Y})$, where $\mathbf{X},\mathbf{Y}$ are assumed to be independent data sets, is given by

$$D(p(\mathbf{X},\mathbf{Y})\|q(\mathbf{X},\mathbf{Y})) = D(p(\mathbf{X})\|q(\mathbf{X})) + D(p(\mathbf{Y})\|q(\mathbf{Y})). \quad (9)$$

IV. STATISTICAL MODELING OF WAVELET SUBBAND COEFFICIENTS VIA SYMMETRIC ALPHA-STABLE DISTRIBUTIONS

In the FE step, an acoustic signal is decomposed into several scales through a multiresolution analysis employing the 1D wavelet transform.[24] The energies of the resulting wavelet coefficients identify the content of the signal at each frequency band (scale). The proposed method is based on the accurate modeling of the tails of the marginal distribution of the wavelet coefficients at each subband: The wavelet subband coefficients in various scales are modeled as $S\alpha S$ random variables.

The $S\alpha S$ distribution is best defined by its characteristic function:[30]

$$\phi(t) = \exp(i\delta t - \gamma^{\alpha}|t|^{\alpha}), \quad (10)$$

where α is the *characteristic exponent*, taking values $0<\alpha\leqslant 2$, δ $(-\infty<\delta<\infty)$ is the *location parameter*, and γ $(\gamma>0)$ is the *dispersion* of the distribution. The characteristic exponent is a shape parameter, which controls the "thickness" of the tails of the density function. The smaller the α, the heavier the tails of the $S\alpha S$ density function. The dispersion parameter determines the spread of the distribution around its location parameter, similar to the variance of the Gaussian. A $S\alpha S$ distribution is called *standard* if $\delta=0$ and $\gamma=1$. The notation $X\sim f_{\alpha}(\gamma,\delta)$ means that the random variable X follows a $S\alpha S$ distribution with parameters α,γ,δ.

In general, no closed-form expressions exist for most $S\alpha S$ density and distribution functions. Two important special cases of $S\alpha S$ densities with closed-form expressions are the Gaussian $(\alpha=2)$ and the Cauchy $(\alpha=1)$. Unlike the Gaussian density which has exponential tails, stable densities

have tails following an algebraic rate of decay ($P(X>x) \sim Cx^{-\alpha}$, as $x \to \infty$, where C is a constant depending on the model parameters), hence random variables following $S\alpha S$ distributions with small α values are highly impulsive.

An important characteristic of non-Gaussian $S\alpha S$ distributions is the nonexistence of second-order moments. Instead, all moments of order p less than α do exist and are called the *fractional lower-order moments*. In particular, the FLOM's of a $S\alpha S$ random variable $X \sim f_\alpha(\gamma, \delta=0)$, are given by:[25]

$$E\{|X|^p\} = (C(p,\alpha)\cdot\gamma)^p, \quad 0 < p < \alpha, \tag{11}$$

where

$$(C(p,\alpha))^p = \frac{2^{p+1}\Gamma\left(\frac{p+1}{2}\right)\Gamma\left(-\frac{p}{\alpha}\right)}{\alpha\sqrt{\pi}\Gamma\left(-\frac{p}{2}\right)} = \frac{\Gamma\left(1-\frac{p}{\alpha}\right)}{\cos\left(\frac{\pi}{2}p\right)\Gamma(1-p)}. \tag{12}$$

During the FE step, the $S\alpha S$ model parameters (α, γ) are estimated using the consistent ML method described by Nolan,[31] which gives reliable estimates and provides the tightest confidence intervals.

The 1D orthogonal discrete wavelet transform (DWT) expands a signal using a certain basis, whose elements are scaled and translated versions of a single prototype filter ("mother wavelet"). In particular, at the first stage the DWT decomposes a signal in one low-frequency *approximation* subband and a high-frequency *detail* subband. The decomposition process can then be iterated, with successive approximation subbands being decomposed in turn, so that one signal is broken down into many lower resolution components (dyadic scales). Thus, a signal decomposed in N levels results in $N+1$ subbands, 1 approximation subband and N detail subbands.

There are interesting properties of the wavelet transform[23] that justify its use in the proposed classification system, among them are the following: *Locality* (signal content can be analyzed in a local area), *multiresolution* (signal is decomposed at a nested set of dyadic scales). Because of these properties, the wavelet transforms of impulsive signals tend to be sparse, resulting in a large number of small magnitude coefficients and a small number of large magnitude coefficients. Importantly, this property is in conflict with the Gaussian assumption, giving rise to peaky and heavy-tailed *non-Gaussian* marginal distributions of the wavelet subband coefficients.

In the proposed data modeling, the statistical fitting proceeds in two steps: First, we assess whether the data deviate from the normal distribution and we determine if they have heavy tails by employing normal probability plots.[32] Then, we check if the data is in the stable domain of attraction by estimating the characteristic exponent α directly from the data and by providing the related confidence intervals. As a further stability diagnostics, we employ the amplitude probability density (APD) curves ($P(|X|>x)$) that give a good indication of whether the $S\alpha S$ fit matches the data near the mode and on the tails of the distribution.

TABLE I. The shallow water environment.

Water depth (H)	200 m
Range (R)	5 km
Central frequency (f_0)	100 Hz
Bandwidth (Δf)	40 Hz
Source/receiver depth	100 m
Sound speed profile in the water:	
$c_w(0)$	1500 m/s
$c_w(\min)$	1490 m/s
$c_w(H)$	1515 m/s
d [depth of min $c_w(z)$]	50 m
Semi-infinite substrate:	
c_{sb}	1600 m/s
ρ_{sb}	1200 kg/m^3

In the subsequent analysis, the effectiveness of a $S\alpha S$ density function for the approximation of the empirical density of the subband coefficients, near the mode and on the tails, is assessed using the signal of Fig. 4(a), which has been simulated using the environmental parameters of Table I. A description of the sea environment corresponding to these environmental parameters will be given in Sec. V.

Figure 1 compares the $S\alpha S$ and generalized Gaussian density (GGD) fits together with the empirical curve for the subbands of the simulated acoustic signal, which was decomposed in three levels using Daubechies' 4 (db4) wavelet.[33] Clearly, the $S\alpha S$ density is superior to the GGD, following more closely both the mode and the tail of the empirical APDs (that is, the APDs where $P(|X|>x)$ is computed as the percentage of the wavelet coefficients with amplitude greater than x, with x varying between 0 and a maximum value

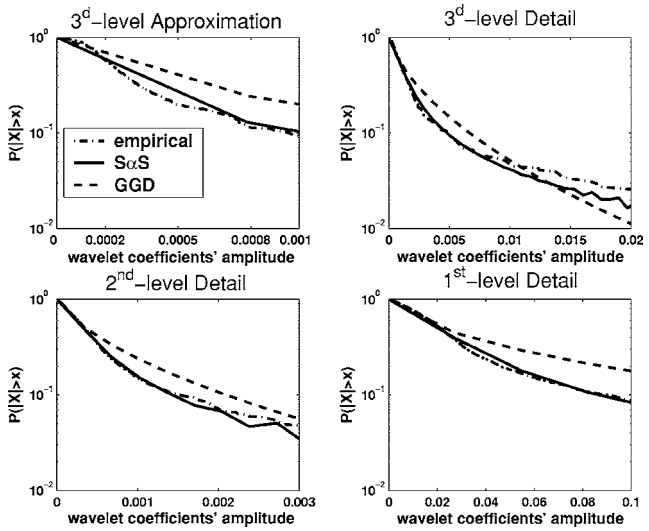


FIG. 1. Modeling of the four wavelet subbands of the three-level decomposition of the *Signal-1* with the $S\alpha S$ and the GGD APD's depicted in solid and dashed lines, respectively. The dashed-dotted line denotes the empirical APD.

TABLE II. $S\alpha S$ modeling of detail subband coefficients of 10 time-domain signals, $\{TD_l\}_{l=20:5:65}$, simulated using the environment of Table I, with $d=l$ m, using Daubechies' 4 wavelet and three decomposition levels. ML parameter estimates and 95% confidence intervals for the characteristic exponent α are given.

	Detail wavelet subbands		
Signal	Level 1	Level 2	Level 3
TD_{20}	1.1792±0.046	1.2104±0.073	1.2119±0.101
TD_{25}	1.1854±0.047	1.2017±0.073	1.1914±0.100
TD_{30}	1.1891±0.048	1.1959±0.075	1.1604±0.102
TD_{35}	1.199±0.048	1.2211±0.074	1.1864±0.102
TD_{40}	1.2031±0.047	1.2159±0.074	1.2294±0.109
TD_{45}	1.2011±0.047	1.1974±0.075	1.1897±0.106
TD_{50}	1.2056±0.046	1.2024±0.075	1.2054±0.104
TD_{55}	1.1978±0.046	1.1978±0.074	1.2503±0.105
TD_{60}	1.2017±0.047	1.2043±0.074	1.2267±0.104
TD_{65}	1.2086±0.047	1.2053±0.073	1.2164±0.103

determined as the maximum amplitude of the available set of wavelet coefficients), than the exponentially decaying GGD. Table II shows the ML estimates of the characteristic exponent α together with the corresponding 95% confidence intervals, for a set of ten simulated acoustic signals obtained using the environment shown in Table I, with d ranging from 20 m to 65 m in steps of 5 m ($d=20:5:65$ m).

Each of them is decomposed in three levels using db4 filters. It can be observed that the confidence intervals depend on the decomposition level. In particular, they become wider as the level increases since the number of samples used for estimating the $S\alpha S$ parameters gets smaller because of the subsampling that takes place between scales. Table II also demonstrates that the coefficients of different subbands exhibit various degrees of non-Gaussianity. Figure 2 displays the histograms of the estimated characteristic exponent values for the simulated acoustic signals in the database to be described in Sec. V, which are also decomposed in three levels using the db4 wavelet, resulting in four wavelet subbands. It is observed that for this set of signals, the wavelet coefficients follow statistics that are highly non-Gaussian and heavy tailed, since the values of the characteristic exponent range in the interval [1,1.4] for each of the subbands. On the other hand, the dispersion γ, with respect to the whole database, ranges in the interval $[10^{-4}, 1.33]$.

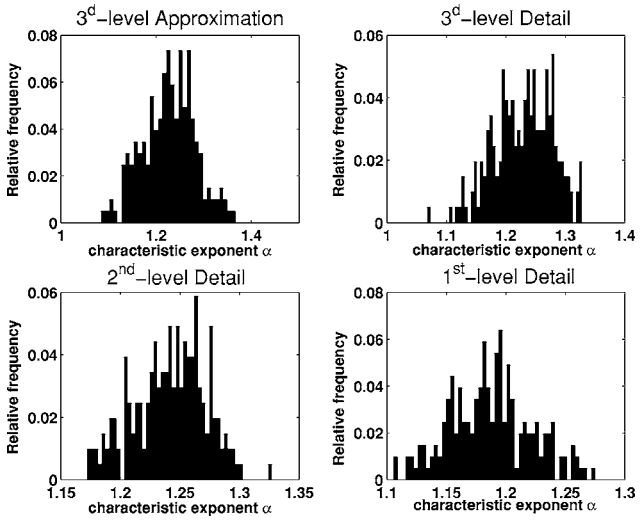


FIG. 2. Histograms representing the relative frequencies of the estimated characteristic exponent values for the 1207 simulated acoustic signals in our database (cf. Sec. V).

A. Feature extraction

After the implementation of the 1D wavelet transform, the marginal statistics of the coefficients at each decomposition level are modeled via a $S\alpha S$ distribution. Then, to extract the features, we simply estimate the (α, γ) pairs at each subband.

Thus, for a given acoustic signal S, decomposed in L levels, its signature is given by the set of the $L+1$ pairs of the estimated parameters:

$$S \mapsto \{(\alpha_1, \gamma_1), (\alpha_2, \gamma_2), \ldots, (\alpha_{L+1}, \gamma_{L+1})\}, \tag{13}$$

where (α_i, γ_i) are the estimated model parameters of the i-th subband. Note that we follow the convention that $i=1$ corresponds to the Detail subband at the first decomposition level, while $i=L+1$ corresponds to the Approximation subband at the L-th level. The total size of the above signature equals $2(L+1)$ which means that the content of an acoustic signal can be represented by only a few parameters, in contrast with the large number of the transform coefficients.

B. Similarity measurement

In the proposed classification scheme, the similarity measurement between two distinct acoustic signals was carried out by employing the KLD. Unfortunately, there is no closed-form expression for the KLD between two general $S\alpha S$ distributions which are not Cauchy or Gaussian. To address this problem, numerical methods could be employed for the computation of the KLD between two numerically approximated $S\alpha S$ densities, resulting in an increased computational burden.

In order to avoid the increased computational complexity of a numerical scheme, first the corresponding characteristic functions are transformed into valid probability density

functions and then the KLD is applied on these normalized versions of the characteristic functions. Due to the one-to-one correspondence between a $S\alpha S$ density and its associated characteristic function, it is expected that the KLD between normalized characteristic functions will be a good similarity measure between the acoustic signals.

If $\phi(\omega)$ is a characteristic function corresponding to a $S\alpha S$ distribution, then the function

$$\hat{\phi}(\omega) = \frac{\phi(\omega)}{c} \tag{14}$$

is a valid density function when

$$c = \int_{-\infty}^{\infty} \phi(\omega) d\omega.$$

For the parameterization of the $S\alpha S$ characteristic function given by Eq. (10), and assuming that the densities are centered at zero, that is $\delta=0$, which is true in the case of wavelet subband coefficients since the average value of a wavelet is zero, the normalization factor is given by

$$c = \frac{2\Gamma\left(\frac{1}{\alpha}\right)}{\alpha\gamma}. \tag{15}$$

By employing the KLD between a pair of normalized $S\alpha S$ characteristic functions, the following closed form expression[27] is obtained:

$$D(\hat{\phi}_1 \| \hat{\phi}_2) = \ln\left(\frac{c_2}{c_1}\right) - \frac{1}{\alpha_1} + \left(\frac{\gamma_2}{\gamma_1}\right)^{\alpha_2} \cdot \frac{\Gamma\left(\frac{\alpha_2+1}{\alpha_1}\right)}{\Gamma\left(\frac{1}{\alpha_1}\right)}, \tag{16}$$

where (α_i, γ_i) are the estimated parameters of the characteristic function $\phi_i(\cdot)$ and c_i is its normalizing factor. It can be proven that $D(\hat{\phi}_1 \| \hat{\phi}_2) \geq 0$ with equality if and only if $(\alpha_1, \gamma_1) = (\alpha_2, \gamma_2)$. That is, the KLD between two signals is minimized when their $S\alpha S$ model parameters are equal.

Thus, the implementation of an L-level DWT on each underwater acoustic signal, results in its representation by $L+1$ subbands, $(D_1, D_2, \ldots, D_L, A_L)$, where D_i, A_i denote the ith level detail and approximation subband coefficients, respectively. Assuming that the wavelet coefficients belonging to different subbands are independent, Eq. (9) yields the following expression for the overall distance between two acoustic signals S_1, S_2:

$$D(S_1 \| S_2) = \sum_{k=1}^{L+1} D(\hat{\phi}_{S_1,k} \| \hat{\phi}_{S_2,k}). \tag{17}$$

V. CLASSIFICATION RESULTS USING SYNTHETIC DATA

In this section, the efficiency of the proposed classification scheme for shallow water acoustic transmissions is evaluated using simulated signals, based on the environment described in Table I. Figure 3 shows the sea environment of the experimental setup, which is assumed range independent and axially symmetric. It consists of a shallow water layer and a semi-infinite bottom (the substrate), which are considered fluid and the sound speed profile may vary with depth in the water layer, while it is constant in the substrate. For simplicity, the density of both layers is assumed to be constant.

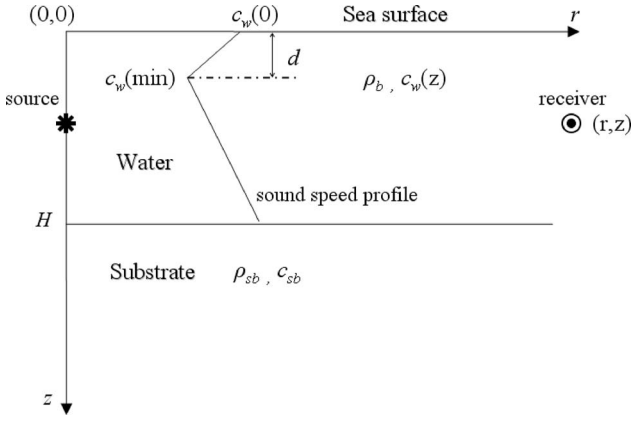


FIG. 3. The shallow water sea environment used in the present study.

The tomographic signals are modeled in the frequency domain using a Gaussian pulse of central frequency $f_0 = 100$ Hz and a bandwidth $\Delta f = 40$ Hz placed at the depth of 100 m. The environmental parameters appear in Table I. First, we generate a set of synthetic signals in order to study the sensitivity of the proposed scheme for small variations of the sound speed in the water column. In particular, for the parameter d, denoting the depth of the minimum sound speed in the water column, a variation in the range $[20, 100]$ m with a step size of 5 m is considered, while the sound speed on the surface varies in the interval $[1495, 1505]$ m/s with a step size equal to 2 m/s and the minimum sound speed at the depth of d m takes values in the interval $[1480, 1500]$ m/s with a step size of 2 m/s. By taking all the permissible combinations of the above parameters according to the sound speed profile shown in Fig. 3, a database with a total of 1207 acoustic signals is obtained, which is equivalent to considering 1207 different sea environments, due to our simulation setup, in which we generate one signal for each setting of the sea parameters. Note that in an actual inversion procedure, the database is never constructed a priori, as more sophisticated algorithms controlling the change of the candidate parameters are applied. However, for the objective of the present study, the construction of the data base using predefined values of the possible environmental parameters is the most appropriate procedure.

The simulated data correspond to measurements at the depth of 100 m and range of 5 km from the source. The data are calculated using the normal-mode program MODE1 developed at FO.R.T.H. These data are provided as input to the inverse discrete Fourier transform to yield the signals in the time domain.

Each of the time-domain signals is decomposed by implementing a three-level 1D DWT using the db2 and db4 wavelets, which have shown the best performance among the wavelets tested for this application. Before proceeding, the APD curves corresponding to the the 4 wavelet subbands of

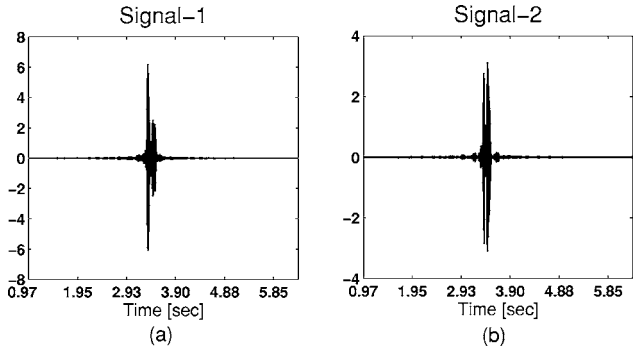


FIG. 4. Acoustic signals simulated in the shallow water environment of Table I with sound speed profiles: (a) $\{c_w(0)=1500\ \mathrm{m/s},\ c_w(\mathrm{min})=1490\ \mathrm{m/s},\ d=50\ \mathrm{m}\}$ and (b) $\{c_w(0)=1500\ \mathrm{m/s},\ c_w(\mathrm{min})=1495\ \mathrm{m/s},\ d=50\ \mathrm{m}\}$.

the two signals "Signal-1" and "Signal-2," shown in Fig. 4, are compared (Fig. 5). Note that Signal-1 corresponds to the parameters of Table I, whereas Signal-2 corresponds to $c_w(\mathrm{min})=1495$ m/s, that is, to a small variation of the minimum sound speed in the water column. It is observed that, although the two signals only slightly differ with the minimum sound speed, their corresponding APDs are clearly distinguishable, especially for the approximation subband and the second-level detail subband, verifying the sensitivity of the $S\alpha S$ model to small variations of the environmental parameters.

In the sequel, the query is the Signal-1. We evaluate the performance of the classification scheme which employs the signatures $\{(\alpha_1,\gamma_1),\ldots,(\alpha_8,\gamma_8)\}$ containing the estimated $S\alpha S$ parameters of the four subbands, as well as the reduced signatures $\{(\alpha_1,\gamma_1),\ldots,(\alpha_6,\gamma_6)\}$ containing the estimated model parameters of the three detail subbands only. The second signature is justified by the fact that the approximation subband, which is a low-pass residual of the original signal, may not preserve the sparsity property of the wavelet transform mentioned above, which results in estimated characteristic exponent values very close or equal to 2, that is the marginal distribution of the approximation subband coefficients is Gaussian.

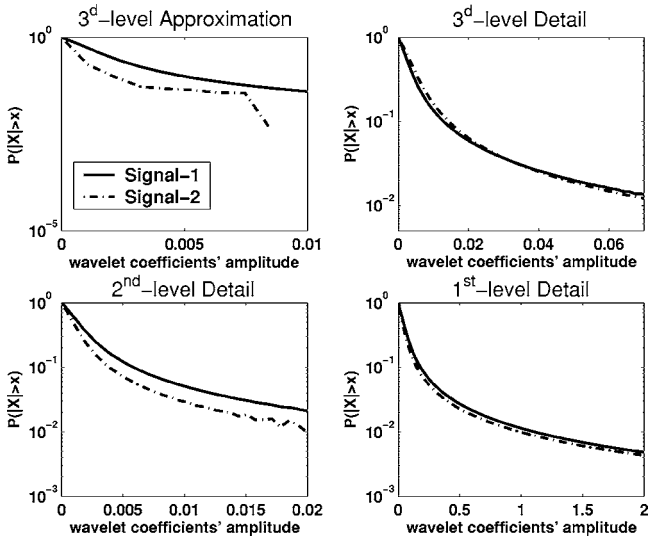


FIG. 5. Comparison between the APD curves of the 4 wavelet subbands for the Signal-1 depicted in solid line and the Signal-2 depicted in dash-dotted line.

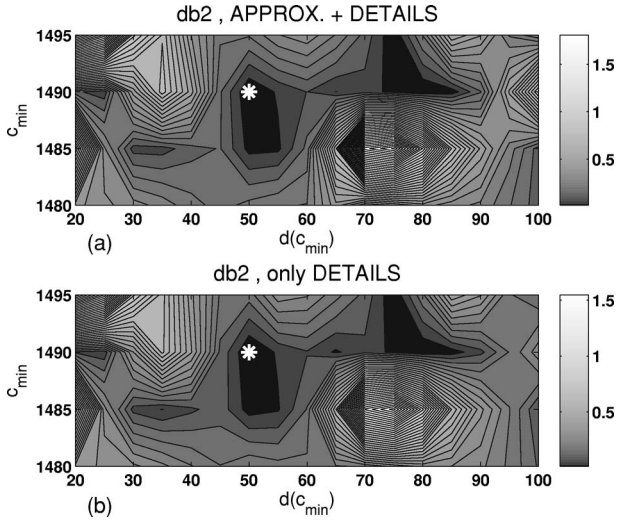


FIG. 6. KLD between Signal-1 and the database signals, decomposed in 3 levels with db2, as a function of $c_w(\mathrm{min})$ and d using (a) all wavelet subbands and (b) only the details.

Figure 6 displays the KLD between Signal-1 and each signal in the database, as a function of the environmental parameters $c_w(\mathrm{min})$ and the depth d, where the sound speed takes its minimum value. All of the signals are decomposed using the db2 wavelet. It is observed that in this case, the use of the two signatures—the original [Fig. 6(a)] and the reduced [Fig. 6(b)]—does not change the regions where the minimum KLD is achieved. The star in the two plots corresponds to zero KLD, that is, its coordinates are equal to the environmental parameters of Signal-1 and this point is unique according to what it has already been mentioned about the KLD between $S\alpha S$ characteristic functions.

It can also be seen that the inclusion of the approximation subband only affects the discrimination power of the KLD between the query signal and the signals which are already "far" from it. Note that the maximum KLD occurs in the same region of the two plots, but it takes a higher value (1.8) when the signature contains the $S\alpha S$ parameters of the approximation subband. On the other hand, taking into account this subband during the similarity measurement step, does not benefit the discrimination power of the KLD in the regions, where it is minimized using only the detail subbands. Besides, this wavelet seems to be unsuitable for an efficient classification scheme, since the regions of minimum KLD are quite large and cover different ranges of the two environmental parameters.

Figure 7 presents the KLD between Signal-1 and each signal in the database, as a function of the same two environmental parameters, $c_w(\mathrm{min})$ and d, when all the signals are decomposed using the db4 wavelet. The improvement, in comparison with the results provided by the db2 wavelet, is obvious. The regions of minimum KLD have been significantly shrunk around the point of zero KLD, denoted by the star, and close to it.

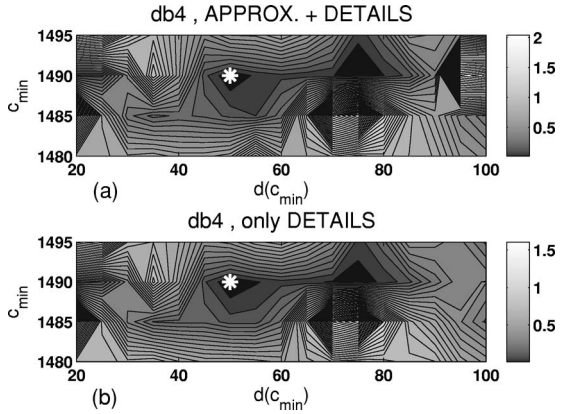


FIG. 7. KLD between Signal-1 and the database signals, decomposed in 3 levels with db4, as a function of c_w(min) and d using (a) all wavelet subbands and (b) only the Details.

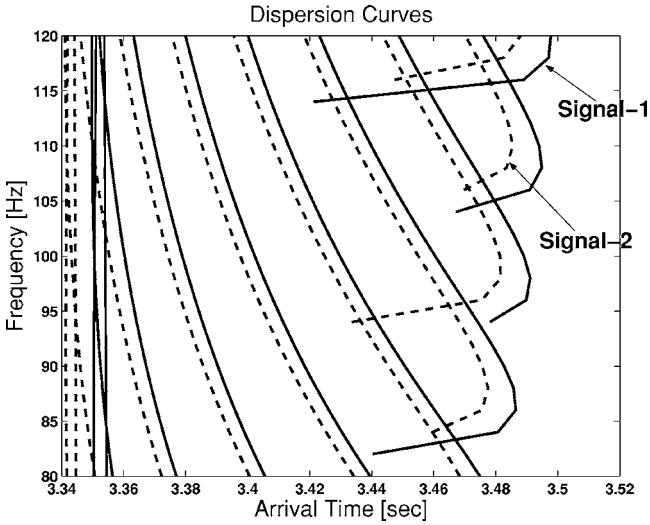


FIG. 8. Dispersion curves corresponding to the simulated signals Signal-1 and Signal-2.

Regarding the comparison with respect to the two signatures, it is observed that the original signature [Fig. 7(a)], containing the $S\alpha S$ parameters of the four subbands, results in a KLD with an increased discrimination power, than the reduced signature [Fig. 7(b)] with the estimated parameters of the detail subbands only. Note also that the maximum KLD (≈ 2) achieved for the original signature, is greater than the maximum KLD corresponding to the reduced one, as in the db2 case. Thus, in the db4 case, the inclusion of the approximation subband is important for the design of an efficient classification scheme.

In order to obtain some indication on the possible advantages of the proposed method for signal classification with respect to other existing ones, we include Fig. 8 which presents the dispersion curves for the two signals under consideration. For shallow water environments, modal inversion schemes are considered effective for the characterization of the environment. For the case under consideration it is clear that lower-order modes, which play an important role in tomographic inversions for the water column, are not well identified for both signals. Although higher-order modes are well identified and they are different for the two signals, no definite conclusion, as regards the efficiency of an inversion scheme based on modal dispersion curves (or modal travel time), can be derived without further analysis. On the other hand, the proposed classification does not suffer from any restrictions as regards the specific character of the signal feature, and therefore it can be considered as a good alternative to signal classification using modal issues, such as modal phase and modal travel time.

As a second illustration, the sensitivity of the $S\alpha S$ model parameters to small variations of the sound speed in the substrate (c_{sb}), as well as the substrate density (ρ_{sb}) is studied. For this purpose, a second synthetic experiment is performed, by generating a set of simulated signals according to the shallow water environmental parameters shown in Table I. The sound speed in the substrate varies in the interval $[1550, 1650]$ m/s with a step size equal to 5 m/s and the substrate density varies in the interval $[1170, 1240]$ kg/m^3 with a step size equal to 1 kg/m^3, resulting in a set with a total of 1491 synthetic acoustic signals.

Each of the above signals is decomposed by implementing a three-level 1D DWT. We tested different wavelet functions, but we present the results only for the db4 wavelet, that gave the best performance. Table III shows the ML es-

TABLE III. $S\alpha S$ modeling of Detail subband coefficients of ten time-domain signals, $\{TD_{\rho_{sb}}\}_{\rho_{sb}=1180:5:1225}$, simulated using the environment of Table I, with substrate density ρ_{sb}, using Daubechies' 4 filter and 3 decomposition levels. ML parameter estimates for the $(\alpha, \gamma/10^{-4})$ parameter pairs are given.

	Detail wavelet subbands		
Signal	Level 1	Level 2	Level 3
TD_{1180}	(1.1576, 10.564)	(1.2058, 3.426)	(1.2531, 211.73)
TD_{1185}	(1.1589, 10.537)	(1.2095, 3.3433)	(1.2491, 211.01)
TD_{1190}	(1.1604, 10.506)	(1.2155, 3.3493)	(1.2454, 210.39)
TD_{1195}	(1.1667, 10.455)	(1.2088, 3.3264)	(1.2428, 209.58)
TD_{1200}	(1.1688, 10.44)	(1.2147, 3.3247)	(1.2397, 208.85)
TD_{1205}	(1.1689, 10424)	(1.2178, 3.3153)	(1.2381, 208.04)
TD_{1210}	(1.1705, 10.472)	(1.2186, 3.3084)	(1.2351, 207.33)
TD_{1215}	(1.1719, 10.408)	(1.2188, 3.2973)	(1.2326, 206.56)
TD_{1220}	(1.1709, 10.39)	(1.2228, 3.2965)	(1.2328, 205.95)
TD_{1225}	(1.1695, 10.369)	(1.2262, 3.2967)	(1.2324, 205.43)

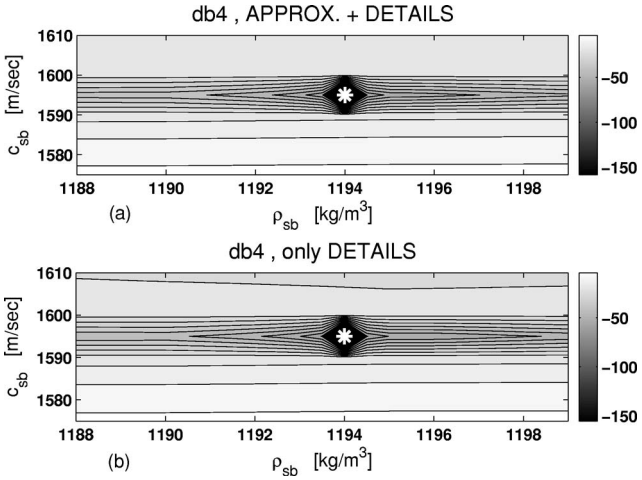


FIG. 9. KLD between the query signal (c_{sb}=1595 m/s, ρ_{sb}=1194 kg/m^3) and the 1491 database signals, decomposed in 3 levels with db4, as a function of c_{sb} and ρ_{sb} using (a) all wavelet subbands and (b) only the details.

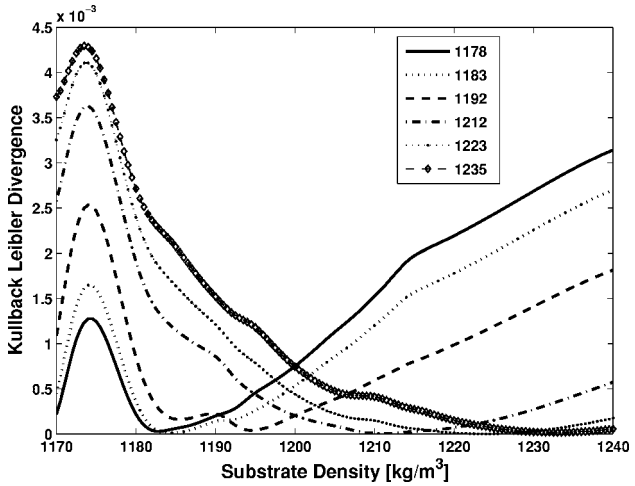


FIG. 10. KLD between each one of the 6 signals, with the same c_{sb} =1600 m/s and ρ_{sb}=1178, 1183, 1192, 1212, 1223 and 1235 kg/m^3, respectively, and the 71 database signals obtained for c_{sb}=1600 m/s and by varying the sediment density in the interval [1170:1:1240] kg/m^3.

timates for the $S\alpha S$ parameter pairs $(\alpha, \gamma/10^{-4})$ for ten signals of the above dataset, with the same sound speed in the substrate (c_{sb}=1600 m/s) and different substrate density taking the values ρ_b=1180:5:1225. It is clear that for these selected signals, the corresponding estimated characteristic exponents α, as well as the estimated dispersions γ are quite close. Thus, one could claim that it would be difficult to classify a given query signal in the correct environment, represented by the database signal whose feature vector is closest to that of the query signal than the feature vectors of the other database signals, since the model parameters corresponding to the wavelet subbands of two different database signals are very close to each other.

However, Fig. 9 displays the KLD between the query, which is the signal corresponding to c_{sb}=1595 m/s and ρ_{sb} =1194 kg/m^3, and each one of the 1491 signals in the database, as a function of the environmental parameters c_{sb} and ρ_{sb}. The values of the KLD are shown in the dB scale and the x,y axes have been properly adjusted, for a better visualization. The region around the location of minimum KLD, denoted by the star in the two plots, is quite tight, indicating a high discrimination power of the KLD between $S\alpha S$ distributions, even in the cases of very small variations of the substrate sound speed and the substrate density. Besides, it is clear that in this case, the use of the original signature [Fig. 9(a)] results in an increased performance compared with the reduced one [Fig. 9(b)].

Figure 10 shows that the proposed classification scheme is able to perform a satisfactory correspondence of a given query to the correct environment, when the substrate density is variable, even in the case that the $S\alpha S$ model parameters are close enough. In particular, each curve in this figure corresponds to the KLD values between each one of the six signals with the same c_{sb}=1600 m/s and different densities ρ_b=1178, 1183, 1192, 1212, 1223, and 1235 kg/m^3, respectively, and the 71 database signals corresponding to c_{sb} =1600 m/s and $\rho_{sb} \in$ [1170:1:1240] kg/m^3. We observe that, for each of the last five signals, the minimum KLD is achieved very close to the corresponding true substrate density, despite the fact that the flatness of the region around the expected density is different for each curve. For the first signal (ρ_{sb}=1178), the difference between the true density and the density that achieves the minimum KLD is slightly larger, but it still remains in a small region on the "substrate density" axis, which is close enough to the true density.

Taking into account the well-known fact, that the density value of the substrate is very difficult to be determined accurately using a typical inversion scheme, this result reinforces our persuasion that a statistical similarity function, such as the KLD, is often preferable than a deterministic one, such as a norm-based distance function between the feature vectors. In addition, it reveals the necessity for the construction of an appropriate similarity function that will be capable of improving the classification performance, even in the cases that the feature extraction results in features that may not be too distinct. The proposed version of the KLD between $S\alpha S$ distributions, seems to be quite powerful with respect to discriminating two simulated acoustic signals with $S\alpha S$ parameters (α, γ) that are very close to each other.

This sensitivity of the KLD to small variations of the sound speed profile in shallow water and the bottom, as well as to small variations of the substrate density, makes this similarity measure an appropriate tool that can be used as the basis of an inversion procedure for geoacoustic inversions.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, the design of a classification scheme for acoustic signals recorded in shallow water, based on a $S\alpha S$ modeling of the coefficients of a 1D wavelet decomposition has been discussed. In particular, it was demonstrated that the set of the parameters of the $S\alpha S$ distributions, describing the statistical behavior of the 1D wavelet subband coefficients, is an effective set of features that can be used for the classification of an underwater acoustic signal. Regarding the task of similarity measurement, the choice of the KLD between normalized $S\alpha S$ characteristic functions, which is a statistical measure of similarity, seems to be appropriate in

capturing the statistical difference of two distinct signals generated even in very similar environments, and thus this KLD is capable in distinguishing two shallow-water environments with approximately the same parameters. This means that the proposed technique can classify, with high probability, an underwater signal in the correct environment where it was recorded. Moreover, it has been shown that the choice of the suitable wavelet function is important for an increased classification performance.

The main computational cost of the proposed scheme, when applied for an inversion problem, is due to the preprocessing step, which includes simulation of a signal in the time-domain using the inverse discrete Fourier transform, followed by the 1D DWT. After this process, the proposed scheme is efficient in terms of computational complexity, since each signal is now represented by only a few parameters, the estimated $S\alpha S$ model parameters. Besides, the KLD has a closed-form expression, which can be simply evaluated using the estimated parameters.

Future research directions, which could further result in an improved classification system with decreased probability of classification error, before a complete inversion procedure is applied for the classification of a sea environment, are the following: First of all, the main assumption throughout the present work was the statistical independence between the wavelet coefficients at adjacent decomposition levels. Regarding the task of similarity measurement between two distinct acoustic signals, we did not take into account the possible interlevel dependencies between the transform coefficients. In this case, we could further improve the power of the similarity measure by considering some kind of chain rule for the KLD between two signals,[22] or by exploiting the possible interdependencies using a multivariate statistical model.

ACKNOWLEDGMENTS

This work was partially supported by the Greek General Secretariat for Research and Technology under Program "THALATTA" and the European Union under Program MC-WAVE.

[1]M. Taroudakis and G. M. Makrakis, (eds.), *Inverse Problems in Underwater Acoustics* (Springer, New York, 2001).
[2]A. Tolstoy, O. Diachok, and L. Frazer, "Acoustic tomography via matched-field processing," J. Acoust. Soc. Am. **89**, pp. 1119–1127 (1991).
[3]A. Tolstoy, *Matched-Field Processing for Underwater Acoustics* (World Scientific, Singapore, 1993).
[4]J. Comput. Acoust. **6**, 1–289 (1998).
[5]S. E. Dosso, M. L. Yeremey, J. M. Ozard, and N. R. Chapman, "Estimation of ocean bottom properties by matched-field inversion of acoustic field data," IEEE J. Ocean. Eng. **18**, 232–293 (1993).
[6]L. Jaschke and N. R. Chapman, "Matched-field inversion of broadband data using the freeze bath method," J. Acoust. Soc. Am. **106**, 1838–1851 (1999).
[7]M. I. Taroudakis and M. G. Markaki, "On the use of matched-field processing and hybrid algorithms for vertical slice tomography," J. Acoust. Soc. Am. **102**, 885–895 (1997).
[8]S. Rajan, J. F. Lynch, and G. V. Frisk, "Perturbative inversion methods for obtaining bottom geoacoustic parameters in shallow water," J. Acoust. Soc. Am. **82**, 998–1017 (1987).
[9]E. C. Shang, "Ocean acoustic tomography based on adiabatic mode theory," J. Acoust. Soc. Am. **85**, 1531–1537 (1989).
[10]M. I. Taroudakis and J. S. Papadakis, "A modal inversion scheme for ocean acoustic tomography," J. Comput. Acoust. **1**, 395–421 (1993).
[11]W. Munk and C. Wunsch, "Ocean acoustic tomography: A scheme for large-scale monitoring," Deep-Sea Res., Part A **26**, 123–161 (1979).
[12]W. Munk, P. Worcester, and C. Wunsch, *Ocean Acoustic Tomography* (Cambridge University Press, Cambridge, UK, 1995).
[13]U. Send *et al.* (The THETIS-2 Group), "Acoustic observations of heat content across the mediterranean sea," Nature (London) **38**, 615–617 (1997).
[14]E. C. Shang and Y. Y. Wang, "On the possibility of monitoring the el niño by using modal ocean acoustic tomography," J. Acoust. Soc. Am. **91**, 136–140 (1992).
[15]M. I. Taroudakis, "A comparison of the modal-phase and modal-travel time approaches for ocean acoustic tomography," in *Proceedings of the 2nd European Conference on Underwater Acoustics*, (edited by L. Bjørnø), pp. 1057–1062 (1994).
[16]M. I. Taroudakis and M. Markaki, "Tomographic inversions in shallow water, using modal travel time measurements," Acta Acust. (Beijing) **87**, 647–658 (2001).
[17]E. L. Lo, J. X. Zhou, and E. C. Shang, "Normal mode filtering in shallow water," J. Acoust. Soc. Am. **74**, 1833–1836 (1983).
[18]M. I. Taroudakis, "Identifying modal arrivals in shallow water for bottom geoacoustic inversions," J. Comput. Acoust. **8**, 307–324 (2000).
[19]G. R. Potty and J. H. Miller, "Geoacoustic tomography: Range dependent inversions on a single slice," J. Comput. Acoust. **8**, 325–346 (2000).
[20]G. R. Potty, J. H. Miller, J. F. Lynch, and K. B. Smith, "Tomographic inversion for sediment parameters in shallow water," J. Acoust. Soc. Am. **108**, 973–986 (2000).
[21]M. I. Taroudakis and G. Tzagkarakis, "On the use of the reassigned wavelet transform for mode identification," J. Comput. Acoust. **12**, 175–196 (2004).
[22]S. Kullback, *Information Theory and Statistics* (Dover, New York, 1997).
[23]S. Mallat, *A Wavelet Tour of Signal Processing* (Academic, New York, 1998).
[24]S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," IEEE Trans. Pattern Anal. Mach. Intell. **11**, 674–692 (1989).
[25]G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance* (Chapman and Hall, New York, 1994).
[26]C. L. Nikias and M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications* (Wiley, New York, 1995).
[27]G. Tzagkarakis and P. Tsakalides, "A statistical approach to texture image retrieval via alpha-stable modeling of wavelet decompositions," *Proceedings of the 5th International Workshop on Image Analysis for Multimedia Interactive Services, Lisboa, Portugal*, April 21–23, 2004.
[28]K. Fukunaga, *Introduction to Statistical Pattern Recognition* (Academic, New York, 1990).
[29]T. Cover and J. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[30]J. P. Nolan, "Parameterizations and modes of stable distributions," Stat. Probab. Lett. **38**, 187–195 (1998).
[31]J. P. Nolan, "Numerical calculation of stable densities and distribution functions," Commun. Stat.-Stochastic Models **13**, 759–774 (1997).
[32]J. Chambers, W. Cleveland, B. Kleiner, and P. Tukey, *Graphical Methods for Data Analysis* (Chapman and Hall, New York, 1983).
[33]C. K. Chui, *An Introduction to Wavelets* (Academic, New York, 1992).

Low-frequency sound scattering by internal waves in the ocean

Alexander G. Voronovich[a] and Vladimir E. Ostashev
NOAA/Earth System Research Laboratory, 325 Broadway, Boulder, Colorado 80305

(Received 21 July 2005; revised 23 December 2005; accepted 27 December 2005)

In this paper, long-range propagation of low-frequency sound through an ocean waveguide with random inhomogeneities in the sound speed is studied. Closed equations for the mean field and correlation function of the sound field are derived using the Chernov method. These equations can be considered as a generalization of equations derived by Dozier and Tappert [J. Acoust. Soc. Am. **63**, 353–365 (1978)], which accounts for 3D effects and cross-modal correlations. The equations derived in this paper in a general form are similar to the equations obtained by many other authors. However, without simplifications these equations are difficult to solve even numerically due to high dimension of the matrices appearing in the equations. Some simplifications of the equations for the mean field and correlation function are suggested that account for narrowness of the angular spectrum of the scattered acoustic field and which make these equations amenable for numerical implementation. To study solutions of the simplified equations, they are additionally averaged along the sound propagation path. This allows us to obtain some analytical results. Using the theory developed, the horizontal coherence length of the sound field is estimated for the GM spectrum of internal waves. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2167058]

PACS number(s): 43.30.Re, 43.30.Bp, 43.30.Qd [AIT] Pages: 1406–1419

I. INTRODUCTION

Studies of low-frequency, long-range sound propagation in the ocean are important for source detection and ranging, communication, acoustic tomography, and other practical problems in underwater acoustics. The amplitudes and phases of sound waves propagating over megameter distances in the ocean exhibit significant fluctuations due to internal waves (IWs) so that their coherence diminishes noticeably. This results in degrading performance of large acoustic arrays. Therefore, studies of the statistical characteristics of sound wave propagating through a fluctuating ocean are of primary importance for the above-mentioned problems.

Since the 1970s, there have been many publications devoted to theoretical studies of low-frequency, long-range sound propagation through a random field of IWs in the ocean. Different theoretical approaches can, in principle, be used to solve this problem, e.g., path integral technique (see Refs. 1 and 2 and references therein) or geometrical acoustics.[3] For low-frequency, long-range sound propagation there is usually only a relatively small number of propagating modes that do not interact with the ocean surface and bottom. Therefore, a modal approach to this problem seems adequate. One of the first papers in which the modal approach was used is by Dozier and Tappert.[4] Neglecting the cross-correlation between different acoustic modes, they considered sound propagation through a random 2D ocean. In particular, Dozier and Tappert obtained a closed set of equations for mean mode intensities. It followed from their theory that, for large propagation distances, different acoustic modes should have the same mean intensities. In Ref. 5, the use of azimuthal harmonics allowed the author to generalize this theory to 3D. However, in our opinion, this approach is not quite convenient since it requires a large number of high-order azimuthal harmonics to describe predominantly forward, narrow-angle scattering. The mean mode intensities were also studied in other papers,[6–8] where a finite number of propagating modes were approximated by a continuum of modes, and some other approximations were used.

The statistical moments of a sound field propagating in a random medium without regular refraction were calculated and studied in detail using the theory of multiple scattering, the diagram technique, the parabolic equation method, and the Markov approximation (e.g., Refs. 9–11). In several papers (Refs. 12 and 13 and references therein), similar approaches were employed to derive formulas for the coherence function of a sound wave propagating through a fluctuating ocean. Within certain approximations, these works took into account both 3D effects and mode coupling. The results obtained were applied to studies of gains of vertical and horizontal arrays located in a fluctuating ocean.[14,15]

In this paper, we reexamine the problem of calculation of the first two statistical moments of low-frequency sound waves propagating over large distances in a fluctuating ocean. Our approach is similar to that used by Dozier and Tappert.[4] We generalize their results by considering sound propagation in a 3D ocean with random inhomogeneities and taking into account both 3D effects and cross-modal correlations. When deriving closed equations for the first two statistical moments of the sound field, we use the Chernov method.[16] This method is based on the assumption that scattering of sound from a slab of the ocean with a horizontal scale of the order of a correlation radius of ocean inhomogeneities is weak. This assumption is valid for sufficiently low acoustic frequencies. Therefore, long-range sound propaga-

[a]Electronic mail: alexander.voronovich@noaa.gov

 0001-4966/2006/119(3)/1406/14/$22.50

tion through the ocean can be considered as a sequence of statistically independent scattering events. Note that the approximations used in the present paper for derivation of closed equations for statistical moments of a sound field are similar to those made in Ref. 17. However, the resulting equations obtained in these papers are different since in Ref. 17 the decomposition of the sound field into acoustic modes was not employed.

Equations for the first two statistical moments of a sound field propagating in a fluctuating ocean presented below in a general form were obtained in Ref. 18. These equations and the approach for their derivation are similar to those in Ref. 19. Furthermore, the equations are similar to equations for the first two statistical moments of a sound field derived by other authors.[12,13,20] In a general form all these equations are very difficult to solve even numerically due to high dimension of the matrices appearing in the equations. Therefore, the main difference between the results obtained in the above-mentioned papers and the present one is in approximations which are employed to simplify the general equations. Here, we propose approximations which conserve the acoustic energy and which make the resulting equations amenable for numerical solution.

The paper is organized as follows. In Sec. II, we introduce a scattering matrix for acoustic modes. Closed equations for the first two statistical moments of the acoustic modes in a general form are obtained in Sec. III, and explicit expressions for the mean sound field and the coherence function of the sound field are presented in Sec. IV. When deriving simplified versions of the general equations, we always ensure that the acoustic energy flux is conserved. The energy conservation law in terms of the description used in this paper is formulated in Sec. V. In Sec. VI, the scattering cross-section matrices are calculated with the use of the first two terms of the Born series. Taking into account the narrowness of the angular spectrum of the scattered acoustic field, simplified versions of the basic equations are derived in Sec. VII. These equations are the main result of the present paper. Before solving them numerically, it is important to ensure that the expected results will make sense. For this reason, in Sec. VIII, these equations are spatially averaged along the sound propagation path. The averaged equations allow us to obtain analytical formulas for the mean sound field and the coherence functions. In Sec. IX, we consider the 2D case and show that after appropriate approximations our equations reduce to those obtained by Dozier and Tappert. Finally, in Sec. X, we calculate the width of the angular spectrum of the acoustic field and show that it is in agreement with the data obtained during the 1998–1999 North Pacific Acoustic Laboratory (NPAL) experiment.[21] Furthermore, we present some preliminary numerical calculations of correlation functions of mode amplitudes.

II. MODE AMPLITUDES AND SCATTERING MATRIX

Let us consider an oceanic waveguide with the sound speed $c(\vec{r},z)$ depending generally on both the vertical z and horizontal $\vec{r}=(x,y)$ coordinates. In this paper, we will consider propagation of a single frequency sound wave. The sound field $p(\vec{r},z)$ of this wave satisfies the Helmholtz equation:

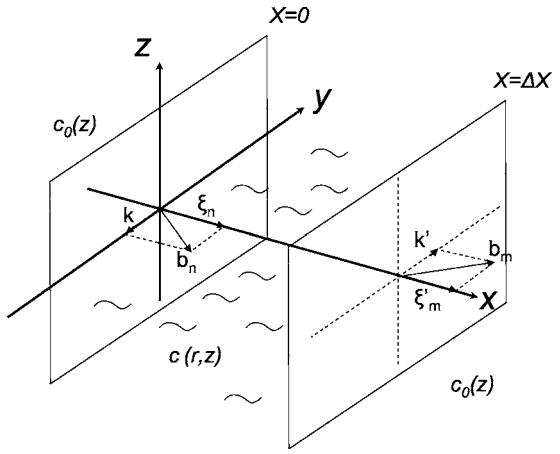


FIG. 1. The geometry of sound scattering.

$$\left(\partial_z^2 + \nabla^2 + \frac{\omega^2}{c^2}\right)p = 0, \tag{1}$$

where $\nabla=(\partial_x,\partial_y)$ represents the horizontal gradient and ω is the sound frequency.

Let us now assume that the waveguide consists of three parts: two horizontally homogeneous parts located at $x<0$ and $x>\Delta x$ where the sound speed $c_0(z)$ depends only on the vertical coordinate z, and the third part being a vertical layer $0<x<\Delta x$ which contains 3D inhomogeneities of the sound speed c (see Fig. 1). The scale of these inhomogeneities is assumed to be much greater than the sound wavelength so that a sound wave is predominantly scattered in the forward direction.

In the homogeneous parts of the wave guide, the Helmholtz equation (1) has solutions in the form of propagating acoustic modes:

$$p_n = u_n(z)\exp[i\xi_n(k)x + iky]. \tag{2}$$

Here, n indicates the acoustic mode index, ξ_n and k are the components of the horizontal wave vector along the x and y axes, respectively, and the mode profiles $u_n(z)$ are solutions of the following boundary problem:

$$\frac{d^2u_n}{dz^2} + \left(\frac{\omega^2}{c_0^2(z)} - b_n^2\right)u_n = 0, \quad u_n(0) = u_n(-\infty) = 0, \tag{3}$$

where b_n are the eigenvalues determined by the sound speed $c_0(z)$ and frequency ω. We will assume that the acoustic mode profiles are normalized as follows:

$$\int_{-\infty}^{0} u_n u_m \, dz = \delta_{nm}, \tag{4}$$

where δ_{nm} is the Kronecker symbol.

The acoustic mode p_n can propagate in an arbitrary direction in the horizontal plane. According to Eqs. (1)–(3), ξ_n and k are related as follows: $\xi_n^2+k^2=b_n^2$. Thus, we have

$$\xi_n = \xi_n(k) = (b_n^2 - k^2)^{1/2}, \quad \mathrm{Im}\,\xi_n \geqslant 0. \tag{5}$$

Note that the direction of horizontal propagation of the nth acoustic mode is uniquely characterized by k.

In the homogeneous parts of the wave guide, an arbitrary acoustic field $p(x,y,z)$ can be represented as a superposition of the modes propagating in all horizontal directions:

$$p(x,y,z) = \sum_n u_n(z) \int dk e^{i\xi_n(k)x + iky} A_n(k), \tag{6}$$

where $A_n(k)$ is a complex amplitude of the nth mode. Note that Eq. (6) describes only sound waves propagating in the positive direction of the x axis since backscattering is neglected throughout this work.

Now let us assume that the nth acoustic mode p_n, given by Eq. (2), is incident from the half space $x<0$ on the inhomogeneous layer of the ocean $0<x<\Delta x$. The sound field at $x>\Delta x$ will be denoted as $p^{(n)}(x,y,z)$. Similar to Eq. (6), this field can be represented in the following form:

$$p^{(n)}(x,y,z) = u_n(z)\exp[i\xi_n(k)x + iky] + \sum_m u_m(z) \int dk' e^{i\xi_m(k')x + ik'y} A_m(k'). \tag{7}$$

The acoustic mode incident on the inhomogeneous layer is written as a first term on the right-hand side of Eq. (7). Note that the complex amplitudes $A_m(k')$ in this equation depend on n and k, and also linearly depend on the amplitude of the incident acoustic mode. Therefore, $A_m(k')$ can be expressed in the following form:

$$A_m(k') = \exp[-i\xi_m(k')\Delta x] S_{mn}(k',k). \tag{8}$$

Here, the exponential term is factored out for convenience, and S_{mn} are the components of the matrix S, which is called the scattering matrix. The explicit expression for this matrix is derived in the Born approximation in Sec. VI. With these notations, the acoustic field at $x>\Delta x$ can be written as

$$p^{(n)} = u_n(z)\exp[i\xi_n(k)x + iky] + \sum_m u_m(z) \int dk' S_{mn}(k',k)\exp[i\xi_m(k')(x-\Delta x) + ik'y]. \tag{9}$$

Description of sound scattering in terms of the scattering matrix S is convenient because it does not depend on the position of sources and/or receivers. The scattering matrix depends only on the inhomogeneities located within the considered inhomogeneous layer and, hence, depends on x and Δx. For convenience, this dependence is not indicated here explicitly. The scattering matrix can be useful if it can be calculated with sufficient accuracy and if the number of propagating modes is not too large. This appears to be the case for low-frequency sound propagation when one can calculate S_{mn} using a perturbative approach.

Let us consider now the acoustic field p in the plane $x=\mathrm{const}$ which can be located within the 3D inhomogeneous layer. This field can be represented as follows:

$$p(x,y,z) = \sum_n u_n(z) \int dk e^{iky} a_n(k,x). \tag{10}$$

Here, $a_n(k,x)$ are the amplitudes of the acoustic modes that depend on x. Note that $a_n(k,x)$ differ from the amplitudes $A_n(k)$ that do not depend on x. A representation of p given by Eq. (10) is unique, since we are considering only waves propagating in the positive direction of the x axis. Using Eqs. (9) and (10), it can be shown that the mode amplitudes at the boundaries of the inhomogeneous layer (i.e., at x and $x+\Delta x$) are related by the following equation:

$$a_n(k,x+\Delta x) = a_n(k,x)e^{i\xi_n(k)\Delta x} + \sum_m \int dk' S_{nm}(k,k') a_m(k',x). \tag{11}$$

Note that this equation can also be derived from the Helmholtz equation (1) represented in the integral form provided that backscattering is neglected. In the absence of scattering ($S_{mn}=0$), using Eqs. (10) and (11), it can be easily shown that the acoustic field p is a sum of noninteracting modes:

$$p(x,y,z) = \sum_n u_n(z) \int dk a_n(k,0) e^{i\xi_n(k)x + iky}. \tag{12}$$

This equation agrees with Eq. (6), as it should, provided that $a_n(k,0)=A_n(k)$.

In the following, we will assume that the ocean inhomogeneities are random with horizontally homogeneous statistics. We will need the first two statistical moments of the scattering matrix. The scattering matrix can always be represented as a sum of its mean value $\bar{S}_{mn}$ and a fluctuating part ΔS_{mn}:

$$S_{mn}(k,k_0) = \Delta x \bar{S}_{mn}(k)\,\delta(k-k_0) + \Delta S_{mn}(k,k_0), \tag{13}$$

The first term on the right-hand side of Eq. (13) is proportional to $\delta(k-k_0)$ due to horizontally homogeneous statistics of random inhomogeneities. (Simple proof of this fact can be found in Sec. 2.4 from Ref. 22.) Furthermore, in this equation Δx is factored out from $\bar{S}_{mn}$ for convenience. Similarly,

$$\langle \Delta S_{n_1 m_1}(k-\eta/2, k_0-\eta_0/2) \Delta S^*_{n_2 m_2}(k+\eta/2, k_0+\eta_0/2)\rangle = \Delta x \mathcal{E}_{n_1 n_2, m_1 m_2}(k,k_0;\eta)\,\delta(\eta-\eta_0). \tag{14}$$

Here, $\mathcal{E}_{n_1 n_2, m_1 m_2}$ are the components of the matrix $\mathcal{E}$, which we call the scattering cross-section matrix.

In the next section, assuming that the matrices $\bar{S}$ and $\mathcal{E}$ are known, we will derive closed equations for the first two statistical moments of mode amplitudes a_n. The matrices $\bar{S}$ and $\mathcal{E}$ will be calculated in Sec. VI.

III. EQUATIONS FOR THE FIRST TWO STATISTICAL MOMENTS OF MODE AMPLITUDES

Let the horizontal correlation radius ρ_{corr} of random inhomogeneities be significantly smaller than Δx and sound scattering by a 3D layer of the ocean with the width of the order of ρ_{corr} be weak in comparison with scattering from the whole layer $(x,x+\Delta x)$. Then, we can neglect statistical de-

pendence between $a_n(x)$ and S_{ij}, where $a_n(x)$ is calculated at point x and S_{ij} is calculated for the 3D layer of the ocean located between x and $x+\Delta x$:

$$\langle a_n S_{ij}\rangle = \langle a_n\rangle\langle S_{ij}\rangle, \tag{15}$$

$$\langle a_n S_{ij} a_{n'}^*, S_{i'j'}^*\rangle = \langle a_n a_{n'}^*\rangle\langle S_{ij} S_{i'j'}^*\rangle. \tag{16}$$

These equations represent the Chernov approximation.[16] Equations (15) and (16) can be justified as follows. In the forward scattering approximation, the value of a_n in the plain $x=$const depends only on the inhomogeneities at $x'<x$. On the other hand, the components S_{ij} of the scattering matrix depend on the inhomogeneities located within the 3D layer $x<x'<x+\Delta x$. A statistical relation between S_{ij} and a_n is due to inhomogeneities located within a 3D layer $x-\rho_{\text{corr}}<x'<x+\rho_{\text{corr}}$. If sound scattering by this layer is weak in comparison to the scattering by the layer $(x, x+\Delta x)$, Eqs. (15) and (16) hold.

Now we substitute Eq. (13) into Eq. (11) and average both sides of the resulting equation. Using the Chernov approximation, Eq. (15), we obtain a closed equation for the mean mode amplitude:

$$\bar{a}_n(k,x+\Delta x) = \bar{a}_n(k,x)e^{i\xi_n(k)\Delta x} + \Delta x\sum_m \bar{S}_{nm}(k)\bar{a}_m(k,x). \tag{17}$$

It follows from this equation that in the Chernov approximation, the mean mode amplitudes with different k do not interact with each other.

To derive an equation for the second moments of the mode amplitudes, we substitute Eq. (13) into Eq. (11) and replace k with $k-\eta/2$:

$$a_n(k-\eta/2,x+\Delta x) = a_n(k-\eta/2,x)e^{i\xi_n(k-\eta/2)\Delta x} + \Delta x\sum_m \bar{S}_{nm}(k-\eta/2)a_m(k-\eta/2,x) + \sum_m \int dk'\,\Delta S_{nm}(k-\eta/2,k')\times a_m(k',x). \tag{18}$$

Similarly,

$$a_{n'}^*(k+\eta/2,x+\Delta x) = a_{n'}^*(k+/2,x)e^{-i\xi_{n'}(k+\eta/2)\Delta x} + \Delta x\sum_{m'} \bar{S}_{n'm'}^*(k+\eta/2)\times a_{m'}^*(k+\eta/2,x) + \sum_{m'} \int dk''\times \Delta S_{n'm'}^*(k+\eta/2,k'')a_{m'}^*(k'',x). \tag{19}$$

Let $B_{nn'}$ be the correlation functions of the mode amplitudes:

$$B_{nn'}(k,x;\eta) = \langle a_n(k-\eta/2,x)a_{n'}^*(k+\eta/2,x)\rangle. \tag{20}$$

The dimension of the correlation functions $B_{nn'}$ is meters. In what follows, if $n=n'$, the correlation functions $B_{nn'}$ will be called mode correlation functions, and if $n\neq n'$, they will be called cross-modal correlation functions.

Substituting Eqs. (18) and (19) into Eq. (20) and using Eqs. (14)–(16), we obtain a closed equation for $B_{nn'}$:

$$\begin{aligned} B_{nn'}(k,x+\Delta x;\eta) = {} & B_{nn'}(k,x;\eta)e^{i[\xi_n(k-\eta/2)-\xi_{n'}(k+/2)]\Delta x} + \Delta x\sum_{m'} e^{i\xi_n(k-\eta/2)\Delta x}\bar{S}_{n'm'}^*(k+\eta/2)B_{nm'}(k,x;\eta) \\ & + \Delta x\sum_m e^{-i\xi_{n'}(k+\eta/2)\Delta x}\bar{S}_{nm}(k-\eta/2)B_{mn'}(k,x;\eta) + \Delta x\sum_{m,m'}\int dk'\,\mathcal{E}_{nn',mm'}(k,k';\eta)B_{mm'}(k',x;\eta) \\ & + (\Delta x)^2\sum_{m,m'} \bar{S}_{nm}(k-\eta/2)\bar{S}_{n'm'}^*(k+\eta/2)B_{mm'}(k,x;\eta). \end{aligned} \tag{21}$$

Note that η in this equation is a parameter; the equations corresponding to different η decouple.

If matrices $\bar{S}$ and $\mathcal{E}$ are known, Eqs. (17) and (21) represent a closed set of equations that allow us to calculate the mean mode amplitudes $\bar{a}_n$ and the correlation functions B_{nm} of mode amplitudes by marching a solution along the x axis with a step Δx. Let us specify initial conditions for Eqs. (17) and (21). Assuming that a point source is located at the depth z_s and $x_s=y_s=0$ and using an appropriate representation of the Green function [see Eq. (38) below], we obtain the following initial conditions:

$$\bar{a}_n(x,k)|_{x=0} = -\frac{i}{4\pi}\frac{u_n(z_s)}{\xi_n(k)}, \tag{22}$$

$$B_{nm}(k,x;\eta)|_{x=0} = \frac{1}{(4\pi)^2}\frac{u_n(z_s)}{\xi_n(k-\eta/2)}\left(\frac{u_m(z_s)}{\xi_m(k+\eta/2)}\right)^*. \tag{23}$$

It follows from Eqs. (22) and (23) that, near the point source, acoustic energy is contained in the coherent part of the sound field. As distance from the source increases, the energy is gradually transformed into the incoherent part of the field. At some distance from the source, the coherent part of the field

becomes very small so that the energy is contained mostly in the incoherent part. At greater distances, sound scattering results in redistribution of the energy between different acoustic modes and different directions of propagation.

Equations (17) and (21) present equations for the first two statistical moments of the mode amplitudes in a general form. In the rest of the paper, we will calculate the matrixes $\bar{S}_{nm}$ and $\mathcal{E}_{nn',mm'}$ appearing in these equations, simplify these equations, and solve the resulting equations. It can be shown that in the limiting case of an ocean without stratification Eqs. (17) and (21) coincide with equations for the first two statistical moments of a sound field propagating in a random medium without regular refraction. The ranges of applicability of the latter two equations are well known:[11]

$$\beta\rho_{\text{corr}} \ll 1, \quad y_* \gg \lambda. \tag{24}$$

Here, β is the extinction coefficient of the mean sound field, y_* is the transverse correlation length of a sound field, and λ is the sound wave length. Since approximations made in derivation of Eqs. (17) and (21) are similar to those in a random medium without regular refraction, we can conclude that inequalities in Eq. (24) are necessary conditions of applicability of Eqs. (17) and (21) if these inequalities are valid for every acoustic mode (β and y_* are considered in detail in Sec. VIII.) Note that β increases and y_* decreases when frequency increases. Therefore, the theory developed is valid for sufficiently low frequencies.

IV. COHERENCE FUNCTION AND MEAN SOUND FIELD

Let $p_1=p(x,y+\Delta y/2,z_1)$ and $p_2=p(x,y-\Delta y/2,z_2)$ be the sound fields at two points of observation located in a plane perpendicular to the x axis. Here, Δy is the distance between these two points along the y axis, and y is their geometrical center along this axis. The second moment $\langle p_1 p_2^* \rangle$ is called the transverse coherence function of the sound field. Using Eq. (10), the transverse coherence function can be written as

$$\langle p_1 p_2^* \rangle = \sum_{n,m} \Gamma_{nm} = \sum_{n,m} u_n(z_1) u_m(z_2) \int dk e^{ik\Delta y} \varphi_{nm}(k,x,y). \tag{25}$$

Here, Γ_{nm} is the coherence function corresponding to the nth and mth modes, and φ_{nm} is the angular spectrum of the correlation function at the point (x,y), given by

$$\varphi_{nm}(k,x,y) = \int d\eta B_{nm}(k,x;\eta) e^{-i\eta y}. \tag{26}$$

The spectrum φ_{nm} can be measured experimentally using a billboard receiving array. Equations (25) and (26) establish a relationship between basic variables of the theory developed in this paper and experimental measurements.

Note that the correlation functions B_{nm} correspond to the total sound field $p=\langle p \rangle + \Delta p$, where $\langle p \rangle$ is the mean sound field and Δp is the fluctuating part of the sound field. The mean sound field can be expressed in terms of the mean mode amplitudes by averaging both sides of Eq. (10):

$$\langle p(x,y,z) \rangle = \sum_n u_n(z) \int dk e^{iky} \bar{a}_n(x,k). \tag{27}$$

The correlation function of the incoherent part of the sound field $\Delta p = p - \langle p \rangle$ can be expressed in terms of the coherence function and the mean sound field:

$$\langle \Delta p_1 \Delta p_2^* \rangle = \langle p_1 p_2^* \rangle - \langle p_1 \rangle \langle p_2^* \rangle. \tag{28}$$

In the absence of scattering,

$$a_n(k,x) \sim \exp(i\xi_n(k)x). \tag{29}$$

For small k, this formula takes the form

$$a_n(k,x) \sim \exp\left[i\left(b_n - \frac{k^2}{2b_n} \right) x \right]. \tag{30}$$

Using Eq. (38) below, it can be shown that the term proportional to k^2 on the right-hand side of Eq. (30) results in the decrease of the mode amplitude a_n due to cylindrical spreading.

Let us average both sides in Eq. (25) over y:

$$\int_{-\infty}^{\infty} \langle p_1 p_2^* \rangle dy = 2\pi \sum_{n,m} u_n(z_1) u_m(z_2) \int dk e^{ik\Delta y} B_{nm}(k,x;0). \tag{31}$$

Thus, to calculate the y-averaged transverse coherence function, we can set $\eta=0$ in Eq. (21).

V. ENERGY CONSERVATION

It can be shown that two solutions $p^{(1)}$ and $p^{(2)}$ of the Helmholtz equation (1) without absorption satisfy the following relationship:

$$\frac{i}{4\pi} \int_x \left(p^{(1)} \frac{\partial}{\partial x} p^{(2)*} - p^{(2)*} \frac{\partial}{\partial x} p^{(1)} \right) dz dy = \frac{i}{4\pi} \int_{x+\Delta x} \left(p^{(1)} \frac{\partial}{\partial x} p^{(2)*} - p^{(2)*} \frac{\partial}{\partial x} p^{(1)} \right) dz dy. \tag{32}$$

Here, the integrals are calculated in the planes $x=\text{const}$ and $x+\Delta x=\text{const}$. Let $p^{(1)}$ and $p^{(2)}$ be given by the right-hand side of Eq. (9) with $k=k_1$, $n=n_1$ and $k=k_2$, $n=n_2$, respectively. Then, from Eq. (32) we obtain the optical theorem:

$$\xi_{n_1} e^{i\xi_{n_1}\Delta x} S^*_{n_1 n_2}(k_1,k_2) + \xi_{n_2} e^{-i\xi_{n_2}\Delta x} S_{n_2 n_1}(k_2,k_1) + \sum_m \int dk' \xi'_m S_{mn_1}(k',k_1) S^*_{mn_2}(k',k_2) = 0. \tag{33}$$

Here, $\xi_{n_1}=\xi_{n_1}(k_1)$, $\xi_{n_2}=\xi_{n_2}(k_2)$, and $\xi'_m=\xi_m(k')$. Substituting Eq. (13) into Eq. (33) and averaging both sides of the resulting equation, we have

$$\xi_{n_1} e^{i\xi_{n_1}\Delta x} \bar{S}^*_{n_1 n_2}(k) + \xi_{n_2} e^{-i\xi_{n_2}\Delta x} \bar{S}_{n_2 n_1}(k) + \sum_m \int dk' \xi'_m \mathcal{E}_{mm,n_1 n_2}(k',k;0) + \Delta x \sum_m \xi_m \bar{S}_{mn_1}(k) \bar{S}^*_{mn_2}(k) = 0. \tag{34}$$

Using this equation, it can be shown that Eq. (21) conserves the energy flux:

$$\sum_n \int dk \xi_n B_{nn}(k,x+\Delta x;0) = \sum_n \int dk \xi_n B_{nn}(k,x;0). \tag{35}$$

VI. CALCULATION OF THE MATRICES $\bar{S}$ AND $\mathcal{E}$

In a stratified ocean without inhomogeneities, the Green function of the Helmholtz equation (1) satisfies the following equation:

$$\left(\partial_z^2 + \nabla^2 + \frac{\omega^2}{c_0^2(z)}\right) G_0(\vec{r};z,z') = \delta(z-z')\delta(\vec{r}). \tag{36}$$

The Green function can be represented in the following form:

$$G_0(\vec{r};z,z') = \sum_m u_m(z)u_m(z') \times \int dk \int dq \frac{1}{(2\pi)^2} \frac{e^{i(qx+ky)}}{b_m^2 - q^2 - k^2}. \tag{37}$$

Integrating over q one finds

$$G_0(\vec{r};z,z') = -\frac{i}{4\pi} \sum_m u_m(z)u_m(z') \times \int \frac{dk}{\xi_m(k)} \exp[i\xi_m(k)x + iky]. \tag{38}$$

In the ocean with 3D inhomogeneities, the use of the Green function G_0 allows us to represent the Helmholtz equation (1) in the integral form:

$$p(\vec{r},z) = p_{in}(\vec{r},z) - \frac{\omega^2}{c_{00}^2} \int \theta(x-x') G_0(\vec{r}-\vec{r}';z,z') \times \Delta n^2(\vec{r}',z') p(\vec{r}',z') d\vec{r}' dz'. \tag{39}$$

Here, p_{in} is the sound field that satisfies the homogeneous equation (36), the step function θ in the integrand is due to the forward-scattering approximation, and Δn^2 is a deviation of the refractive index from its mean value $c_{00}/c_0(z)$:

$$\Delta n^2(\vec{r},z) = \frac{c_{00}^2}{c^2(\vec{r},z)} - \frac{c_{00}^2}{c_0^2(z)}, \tag{40}$$

where c_{00} is a reference sound speed.

Let the incident sound field p_{in} coincide with the first term on the right-hand side of Eq. (9):

$$p_{in} = u_n(z)\exp(i\xi_n(k)x + iky). \tag{41}$$

Solving Eq. (39) by iterations, we obtain the following expression for the singly scattered sound field:

$$p_{sc}^{(1)}(x,y,z) = \frac{i}{4\pi}\frac{\omega^2}{c_{00}^2} \sum_m u_m(z)u_m(z') \int_0^{\Delta x} dx' \int dy'dz' \int \frac{dk'}{\xi_m(k')} \exp[i\xi_m(k')(x-x') + ik'(y-y')] \Delta n^2(\vec{r}',z')u_n(z') \times \exp[i\xi_n(k)x' + iky']. \tag{42}$$

Comparing this formula with Eq. (9), we find an explicit expression for the scattering matrix in the first Born approximation:

$$S_{mn}(k',k) = \frac{i}{4\pi}\frac{\omega^2}{c_{00}^2} \frac{\exp(i\xi_m(k')\Delta x)}{\xi_m(k')} \int_0^{\Delta x} dx \int dy \times \exp[-i(\xi_m(k') - \xi_n(k))x - i(k'-k)y] \times \int dz \Delta n^2(x,y,z)u_m(z)u_n(z). \tag{43}$$

Let the correlation function of the refractive index fluctuations Δn^2 be given by

$$\langle \Delta n^2(\vec{r}_1,z_1)\Delta n^2(\vec{r}_2,z_2)\rangle = N(\vec{r}_1 - \vec{r}_2;z_1,z_2). \tag{44}$$

Introducing a spectrum of Δn^2, we represent correlation function N in the following form:

$$N(\vec{r},z_1,z_2) = \sum_j \int d\vec{q} e^{i\vec{q}\vec{r}} P_j(q_x,q_y)\Phi_j(\vec{q},z_1)\Phi_j(\vec{q},z_2). \tag{45}$$

Here $\Phi_j(\vec{q},z)$ is a mode function, j is a mode number, and P_j is the spectral density that depends on a mode number and horizontal wave vector $\vec{q}$. If the refractive index fluctuations are due to linear IWs, the functions $\Phi_j(\vec{q},z)=\Phi_j(q,z)$ are proportional to the mode profiles of these waves. [Generally, Φ_j are eigenfunctions of the symmetrical positively determined integral operator, with the kernel being a Fourier transform of $N(\vec{r},z_1,z_2)$.]

To calculate the scattering cross-section matrix $\mathcal{E}$, we substitute Eq. (43) into Eq. (14) and replace N with the right-hand side of Eq. (45). The resulting equation contains integrals with respect to x_1 and x_2, both taken from 0 and Δx. Introducing new variables $x=(x_1+x_2)/2$ and $x'=x_1-x_2$ and taking into account that ρ_{corr} is significantly smaller than Δx, we extend the integration over x' to infinity. As a result, we obtain the following expression for the scattering matrix:

$$\mathcal{E}_{nn',mm'}(k,k';\eta)=\frac{\pi}{2}\left(\frac{\omega}{c_{00}}\right)^4\frac{\exp[i(\xi_n(k-\eta/2)-\xi_{n'}(k+\eta/2))\Delta x]}{\xi_n(k-\eta/2)\xi_{n'}(k+\eta/2)}\frac{1-e^{-i\kappa\Delta x}}{i\kappa\Delta x}\sum_j P_j(\Delta\xi,k-k')N^{(j)}_{nm}(q)N^{(j)}_{n'm'}(q). \tag{46}$$

Here

$$\Delta\xi=\tfrac{1}{2}[\xi_n(k-\eta/2)-\xi_m(k'-\eta/2)+\xi_{n'}(k+\eta/2)-\xi_{m'}(k'+\eta/2)], \tag{47}$$

$$\kappa=\xi_n(k-\eta/2)-\xi_m(k'-\eta/2)-\xi_{n'}(k+\eta/2)+\xi_{m'}(k'+\eta/2), \tag{48}$$

$$N^{(j)}_{nm}(q)=\int dz u_n(z)u_m(z)\Phi_j(q,z), \tag{49}$$

and

$$q=\sqrt{(\Delta\xi)^2+(k'-k)^2}. \tag{50}$$

Note that matrix $\mathcal{E}$ is dimensionless. Also note that the wave vectors of the acoustic mode IWs are related by Eq. (47). Therefore, entries of the matrix $\mathcal{E}$ describe resonant interaction between the acoustic and IW modes.

Calculation of $\bar{S}_{nm}$ to the accuracy of $O(P_j)$ requires another iteration of Eq. (39):

$$p^{(2)}_{sc}(\vec{r},z)=-\frac{\omega^2}{c^2_{00}}\int\theta(x-x')G_0(\vec{r}-\vec{r}';z,z')\Delta n^2(\vec{r}',z')\times p^{(1)}_{sc}(\vec{r}',z')d\vec{r}'dz'. \tag{51}$$

After transformations similar to those resulting in Eq. (46), we obtain the following expressions for the components of the matrix $\bar{S}$:

$$\bar{S}_{mn}(k)=-\frac{1}{4}\left(\frac{\omega}{c_{00}}\right)^4\frac{1}{\xi_m(k)}\frac{\exp[i\xi_n(k)\Delta x]-\exp[i\xi_m(k)\Delta x]}{i(\xi_n(k)-\xi_m(k))\Delta x}\sum_{l,j}\int dq_x\frac{dk'}{\xi_l(k')}P_j(q_x,k'-k)N^{(j)}_{ml}(\sqrt{q_x^2+(k'-k)^2})N^{(j)}_{nl}\times(\sqrt{q_x^2+(k'-k)^2})\int_0^\infty\exp[i(\widetilde{\Delta\xi}-q_x)x']dx', \tag{52}$$

where

$$\widetilde{\Delta\xi}=\xi_l(k')-\frac{\xi_n(k)+\xi_m(k)}{2}. \tag{53}$$

Introducing infinitesimally small attenuation in the last exponent in Eq. (52), we find

$$\int_0^\infty\exp[i(\widetilde{\Delta\xi}-q_x)x']dx'=i\cdot\text{v.p.}\left(\frac{1}{\widetilde{\Delta\xi}-q_x}\right)+\pi\delta(\widetilde{\Delta\xi}-q_x). \tag{54}$$

Using this expression in Eq. (52), we obtain

$$\bar{S}_{mn}(k)=-\frac{1}{4}\left(\frac{\omega}{c_{00}}\right)^4\frac{1}{\xi_m(k)}\frac{\exp[i\xi_n(k)\Delta x]-\exp[i\xi_m(k)\Delta x]}{i(\xi_n(k)-\xi_m(k))\Delta x}\sum_{l,j}\int\frac{dk'}{\xi_l(k')}\left[\pi P_j(\widetilde{\Delta\xi},k'-k)N^{(j)}_{ml}(\tilde{q})N^{(j)}_{nl}(\tilde{q})+i\cdot\text{v.p.}\int dq_x\times\frac{P_j(q_x,k'-k)N^{(j)}_{ml}(\sqrt{q_x^2+(k'-k)^2})N^{(j)}_{nl}(\sqrt{q_x^2+(k'-k)^2})}{\widetilde{\Delta\xi}-q_x}\right], \tag{55}$$

where

$$\tilde{q} = \sqrt{(\widetilde{\Delta\xi})^2 + (k' - k)^2}. \tag{56}$$

Substituting Eqs. (46) and (56) into the optical theorem, Eq. (34), we find that the first three terms cancel each other. The fourth term in Eq. (34) should be neglected since it is of the second order with respect to P_j [to cancel this term, one has to calculate $\bar{S}_{mn}$ and $\mathcal{E}_{nn',mm'}$ to the accuracy of $O(P_j^2)$].

VII. SIMPLIFIED EQUATIONS

In this section, Eqs. (17) and (21) are simplified by taking into account that the angular spectrum of the scattered sound field is narrow. According to Eq. (38), the dependence of $e^{i\xi_n(k)\Delta x}$-type factors on k is important for a correct description of cylindrical spreading. On the other hand, due to predominantly forward scattering, the width of the angular spectrum should be rather small (if x is not too large) so that $k \ll b_n$. This allows one to use the approximation $\xi_n(k) \approx b_n$ when evaluating the integrals appearing in the matrices $\bar{S}$ and $\mathcal{E}$.

Using this approximation in Eq. (55), we have

$$\bar{S}_{nm}(k) = \frac{1}{b_n} e^{i\xi_n(k)\Delta x} \theta^{(1)}_{nm}(k) s_{nm}. \tag{57}$$

Here, the function $\theta^{(1)}_{nm}(k)$ is given by

$$\theta^{(1)}_{nm}(k) = \frac{1 - \exp(-it)}{it}, \quad t = [\xi_n(k) - \xi_m(k)]\Delta x. \tag{58}$$

This function has the following symmetries:

$$\theta^{(1)}_{nm}(k) = \theta^{(1)}_{nm}(-k), \quad \theta^{(1)}_{nm}(k) = [\theta^{(1)}_{mn}(k)]^*. \tag{59}$$

Furthermore, in Eq. (57), s_{nm} do not depend on k and are given by

$$s_{nm} = -\frac{1}{4}\left(\frac{\omega}{c_{00}}\right)^4 \sum_{l,j} \frac{1}{b_l} \int dk' \left[\pi P_j(\widetilde{\Delta\xi}, k') N^{(j)}_{ml}(\tilde{q}) N^{(j)}_{nl}(\tilde{q}) + i \cdot \text{v.p.} \int dq_x \frac{P_j(q_x, k') N^{(j)}_{ml}(\sqrt{q_x^2 + k'^2}) N^{(j)}_{nl}(\sqrt{q_x^2 + k'^2})}{\widetilde{\Delta\xi} - q_x} \right], \tag{60}$$

where now

$$\widetilde{\Delta\xi} = b_l - \frac{b_n + b_m}{2} \tag{61}$$

and

$$\tilde{q} = \sqrt{(\widetilde{\Delta\xi})^2 + k'^2} = \sqrt{\left(b_l - \frac{b_n + b_m}{2}\right)^2 + k'^2}. \tag{62}$$

The constant matrix s_{nm} is symmetric:

$$s_{nm} = s_{mn}. \tag{63}$$

Using Eq. (57), Eq. (17) for the mean mode amplitude can be written in the following form:

$$\bar{a}_n(k, x + \Delta x) = e^{i\xi_n(k)\Delta x} \left[\bar{a}_n(k, x) + \frac{\Delta x}{b_n} \sum_m \times \theta^{(1)}_{nm}(k) s_{nm} \bar{a}_m(k, x) \right]. \tag{64}$$

The scattering cross-section matrix $\mathcal{E}$ given by Eq. (46) can be simplified analogously:

$$\mathcal{E}_{nn',mm'}(k, k'; \eta) = \frac{1}{b_n b_{n'}} e^{i[\xi_n(k - \eta/2) - \xi_{n'}(k + \eta/2)]\Delta x} \times \theta^{(2)}_{nn',mm'}(k'; \eta) e_{nn',mm'}(k - k'), \tag{65}$$

where

$$\theta^{(2)}_{nn',mm'}(k; \eta) = \frac{1 - \exp(-it)}{it},$$

$$t = [\xi_n(k - \eta/2) - \xi_{n'}(k + \eta/2) - \xi_m(k - \eta/2) + \xi_{m'}(k + \eta/2)]\Delta x \tag{66}$$

satisfies the relationships:

$$\theta^{(2)}_{nn',mm'}(k; \eta) = \theta^{(2)}_{nn',mm'}(-k; -\eta), \tag{67}$$

$$\theta^{(2)}_{nn',mm'}(k; \eta) = [\theta^{(2)}_{n'n,m'm}(-k; \eta)]^*. \tag{68}$$

Furthermore, in Eq. (65), $e_{nn',mm'}(k)$ is given by

$$e_{nn',mm'}(k) = \frac{\pi}{2}\left(\frac{\omega}{c_{00}}\right)^4 \sum_j P_j(\Delta\xi, k) N^{(j)}_{nm}(q) N^{(j)}_{n'm'}(q), \tag{69}$$

$$\Delta\xi = \tfrac{1}{2}(b_n + b_{n'} - b_m - b_{m'}), \tag{70}$$

$$q = \sqrt{(\Delta\xi)^2 + k^2} = \sqrt{\tfrac{1}{4}(b_n + b_{n'} - b_m - b_{m'})^2 + k^2}. \tag{71}$$

When deriving these equations, we set $k' \approx k$ in Eq. (48). The matrix $e_{nn',mm'}$ has the following symmetries:

$$e_{nn',mm'}(k) = e_{n'n,m'm}(k), \quad e_{nn',mm'}(k) = e_{nn',mm'}(-k). \tag{72}$$

Using Eqs. (60) and (69), it can be shown that the following relationship holds:

$$s_{nm} + s^*_{nm} + \sum_l \frac{1}{b_l} \int e_{ll,nm}(k) dk = 0. \tag{73}$$

This relationship is analogous to the optical theorem.

The use of Eqs. (60) and (65) allows us to simplify Eq. (21) for the correlation functions $B_{nn'}$:

$$B_{nn'}(k,x+\Delta x;\eta)=e^{i[\xi_n(k-\eta/2)-\xi_{n'}(k+\eta/2)]\Delta x}\Bigg\{B_{nn'}(k,x;\eta)+\frac{\Delta x}{b_{n'}}\sum_{m'}\theta^{(1)}_{m'n'}\left(k+\frac{\eta}{2}\right)s^{*}_{n'm'}B_{nm'}(k,x;\eta)+\frac{\Delta x}{b_n}\sum_{m}\theta^{(1)}_{nm}\Bigg(k-\frac{\eta}{2}\Bigg)s_{nm}B_{mn'}(k,x;\eta)+\frac{\Delta x}{b_nb_{n'}}\sum_{m,m'}\int\theta^{(2)}_{nn',mm'}(k';\eta)e_{nn',mm'}(k-k')B_{mm'}(k',x;\eta)dk'\Bigg\}. \tag{74}$$

When deriving this equation, the last term in Eq. (21), which is proportional to $O(P_j^2)$, was neglected.

The approximations used in derivation of Eq. (74) are consistent with the conservation of energy. Indeed, let us set $\eta=0$ and $n=n'$ in Eq. (74). We apply the integral operator $b_n\int dk$ to both sides of this equation and sum the result over n. Then, we have

$$\sum_n b_n\int dkB_{nn}(k,x+\Delta x;0)=\sum_n b_n\int dkB_{nn}(k,x;0)+\Delta x\int dk\sum_{n,m'}\theta^{(1)}_{m'n}(k)s^{*}_{nm'}B_{nm'}(k,x;0)+\Delta x\int dk\sum_{n,m}\theta^{(1)}_{nm}(k)\times s_{nm}B_{mn}(k,x;0)+\Delta x\int dk\sum_{m,m'}B_{mm'}(k,x;0)\sum_l\frac{1}{b_l}\theta^{(2)}_{ll,mm'}(k,0)\int dk'e_{ll,mm'}(k-k'). \tag{75}$$

Using Eqs. (63) and (73) and taking into account that

$$\theta^{(2)}_{ll,mm'}(k;0)=\theta^{(1)}_{m'm}(k), \tag{76}$$

Eq. (75) can be reduced to the following:

$$\sum_n b_n\int dkB_{nn}(k,x;0)=\text{const}. \tag{77}$$

Thus, the sound energy flux is conserved in this approximation.

VIII. SPATIAL AVERAGING

Before attempting to solve Eqs. (64) and (74) numerically, it is worthwhile to try solving these equations analytically for some particular cases. To achieve this goal, Eqs. (64) and (74) are additionally simplified by spatial averaging along the sound propagation path. Note that this simplification may or may not be adequate for concrete applications.

If sound scattering is sufficiently weak, the dependencies of $\bar{a}_n$ and $B_{nn'}$ on x are given primarily by the exponential factors in the first terms on the right-hand side of Eqs. (64) and (74):

$$\bar{a}_n\sim e^{i\xi_n(k)x},\quad B_{nn'}\sim e^{i(\xi_n(k-\eta/2)-\xi_{n'}(k+\eta/2))x}. \tag{78}$$

This allows us to average Eqs. (64) (after extracting the phase factor $\exp[i\xi_n(k)x]$) and (74) with respect to x over a horizontal distance that is large enough so that all oscillating exponential terms describing interference effects disappear, and that is small enough so that the functions $\bar{a}_ne^{-i\xi_n(k)x}$ and $B_{nn'}$ do not change significantly. Similarly, in Eq. (64) only the exponential terms with the same period should be retained. As a result, we have

$$e^{-i\xi_n(k)\Delta x}\bar{a}_n(k,x+\Delta x)=\bar{a}_n(k,x)+\Delta x\frac{s_{nn}}{b_n}\bar{a}_n(k,x), \tag{79}$$

$$e^{-i[\xi_n(k-\eta/2)-\xi_n(k+\eta/2)]\Delta x}B_{nn}(k,x+\Delta x;\eta)=B_{nn}(k,x;\eta)+\frac{\Delta x}{b_n}(s_{nn}+s^{*}_{nn})B_{nn}(k,x;\eta)+\frac{\Delta x}{b_n^2}\sum_m\int dk'\times e_{nn,mm}(k-k')B_{mm}(k',x;\eta). \tag{80}$$

(Now $\bar{a}_n$ and B_{nn} are spatially averaged functions; however, we will not reflect this here by changing notations.) It follows from Eq. (80) that, after spatial averaging, calculation of the mode correlation functions B_{nn} decouples from calculations of the cross-modal correlation functions B_{nm}.

It is worthwhile to replace the finite-difference Eqs. (79) and (80) with ordinary differential equations by formally setting $\Delta x\to 0$:

$$\frac{\partial\bar{a}_n}{\partial x}-i\xi_n(k)\bar{a}_n=\frac{s_{nn}}{b_n}\bar{a}_n, \tag{81}$$

$$\frac{\partial I_n}{\partial x}-\frac{i\eta k}{2b_n}I_n=\frac{2\,\mathrm{Re}\,s_{nn}}{b_n}I_n+\frac{1}{b_n^2}\sum_m\int dk'\sigma_{nm}(k-k')\times I_m(k'). \tag{82}$$

Here

$$I_n=I_n(k,x;\eta)=B_{nn}(k,x;\eta) \tag{83}$$

and

$$\sigma_{nm}(k)=e_{nn,mm}(k). \tag{84}$$

When deriving Eq. (82), we expanded the difference $\xi_n(k-\eta/2)-\xi_{n'}(k+\eta/2)$ into the Taylor series with respect to k and η (recall that k and η are much less than b_n) and retained

only the first-order terms. Using Eq. (69), it can be shown that σ_{nm} is given by

$$\sigma_{nm}(k) = \frac{\pi}{2}\left(\frac{\omega}{c_{00}}\right)^4 \sum_j P_j(b_n - b_m, k) \times [N_{nm}^{(j)}(\sqrt{(b_n - b_m)^2 + k^2})]^2. \quad (85)$$

Apparently, $\sigma_{nm}(k)$ is a positively determined symmetric matrix with positive entries:

$$\sigma_{nm} = \sigma_{mn} > 0. \quad (86)$$

Using Eq. (73), one finds

$$2\,\mathrm{Re}\, s_{nn} = -\sum_m \frac{1}{b_m}\int dk \sigma_{nm}(k). \quad (87)$$

The solution of Eq. (81) can be written as

$$\bar{a}_n(k,x) = \bar{a}_n(k,0)e^{i\xi_n(k)x - \beta_n x}, \quad (88)$$

where β_n is the extinction coefficient of the nth acoustic mode, given by

$$\beta_n = -\frac{2\,\mathrm{Re}\, s_{nn}}{b_n} = \frac{1}{b_n}\sum_m \frac{1}{b_m}\int dk \sigma_{nm}(k) > 0. \quad (89)$$

Substituting Eq. (88) into Eq. (27) and using the initial condition Eq. (22) we obtain the mean sound field:

$$\langle p(x,y,z)\rangle = -\frac{i}{4\pi}\sum_n u_n(z)u_n(z_s)e^{-\beta_n x}\int \frac{dk}{\xi_n(k)}e^{i\xi_n(k)x + iky}, \quad (90)$$

or

$$\langle p(x,y,z)\rangle = -\frac{i}{4}\sum_n u_n(z)u_n(z_s)H_0^{(1)}(b_n\sqrt{x^2+y^2})e^{-\beta_n x}. \quad (91)$$

According to Eq. (31), to calculate the y-averaged coherence function, it is sufficient to consider the case $\eta = 0$. In this case, Eq. (82) becomes

$$b_n\frac{\partial I_n}{\partial x} = -\sum_m \frac{1}{b_m}\int dk' \sigma_{nm}(k')I_n + \frac{1}{b_n}\sum_m \times \int dk' \sigma_{nm}(k-k')I_m(k'). \quad (92)$$

This equation can be solved by using a Fourier transform. However, we will simplify this equation further by taking into account that functions $\sigma_{nm}(k-k')$ decrease rapidly as $|k-k'|$ increases. This allows us to expand $I_m(k')$ into the Taylor series at the point $k' = k$ and keep the two lowest-order terms. As a result, we obtain

$$\frac{\partial \tilde{I}_n}{\partial x} = -\beta_n \tilde{I}_n + \sum_m \sigma_{nm}^{(0)}\tilde{I}_m + \sum_m \sigma_{nm}^{(2)}\frac{\partial^2 \tilde{I}_m}{\partial k^2}. \quad (93)$$

Here

$$\tilde{I}_n = b_n I_n(k,x;0), \quad (94)$$

and

$$\sigma_{nm}^{(0)} = \frac{1}{b_n b_m}\int dk \sigma_{nm}(k), \quad (95)$$

$$\sigma_{nm}^{(2)} = \frac{1}{2b_n b_m}\int dk \sigma_{nm}(k)k^2. \quad (96)$$

Using Eqs. (89) and (95), the extinction coefficient β_n can be expressed in terms of σ_{nm}^0:

$$\beta_n = \sum_m \sigma_{nm}^{(0)}. \quad (97)$$

Equation (93) is a transport equation and, hence, conserves energy:

$$\frac{\partial}{\partial x}\sum_n \int \tilde{I}_n dk = 0. \quad (98)$$

The initial condition for Eq. (93) can be obtained with the use of Eq. (23):

$$\tilde{I}_n|_{x=0} = \frac{1}{(4\pi)^2}\frac{1}{b_n}|u_n(z_s)|^2. \quad (99)$$

The transport equation (93) has constant coefficients and can be solved. We will consider an asymptotic solution of this equation when $x \to \infty$. Let

$$\tilde{I}_n \sim e^{i\mu x + iqk}. \quad (100)$$

Then, from Eq. (93) we have

$$i\mu\tilde{I}_n = -\beta_n\tilde{I}_n + \sum_m \sigma_{nm}^{(0)}\tilde{I}_m - q^2\sum_m \sigma_{nm}^{(2)}\tilde{I}_m. \quad (101)$$

Solution of this generalized eigenvalue problem gives a relationship between μ and q. Here, the unperturbed matrix $\sigma - \mathrm{diag}(\beta)$ has one zero eigenvalue that corresponds to the eigenvector $\tilde{I}_1 = \tilde{I}_2 = \cdots \tilde{I}_n = 1$ since, according to Eq. (97),

$$\sum_n \beta_n = \sum_{m,n} \sigma_{nm}^{(0)}. \quad (102)$$

(Note that this equation is equivalent to energy conservation.) All other eigenvalues of Eq. (101) are negative so that the corresponding solutions are exponentially small for $x \to \infty$. For small q, the values of μ can be found by a perturbation approach. Let $\delta\tilde{I}$ be a perturbed eigenvector. Then, using Eq. (101), we obtain

$$i\mu = -\beta_n\delta\tilde{I}_n + \sum_m \sigma_{nm}^{(0)}\delta\tilde{I}_m - q^2\sum_m \sigma_{nm}^{(2)}. \quad (103)$$

Summing these equations over n and using Eq. (102), one finds

$$\mu = iq^2\frac{1}{N_{\mathrm{mod}}}\sum_{n,m}\sigma_{nm}^{(2)}. \quad (104)$$

Using Eq. (85), the sum on the right-hand side of Eq. (104) can be written as

$$\frac{1}{N_{\mathrm{mod}}}\sum_{n,m}\sigma_{nm}^{(2)} = k_0^3\alpha. \quad (105)$$

Here

$$k_0 = \frac{\omega}{c_{00}}, \tag{106}$$

and

$$\alpha = \frac{\pi}{2}\frac{k_0}{N_{\mathrm{mod}}}\sum_{n,m,j}^{N_{\mathrm{mod}}}\frac{1}{2b_nb_m}\int k^2 P_j(b_n - b_m, k) \times [N_{nm}^{(j)}(\sqrt{(b_n - b_m)^2 + k^2})]^2\, dk. \tag{107}$$

The dimensionless parameter α has a meaning of the variance of the scattering angle after passing a horizontal distance of one sound wavelength. Thus, for large x we obtain the following solution of the transport equation (93):

$$\tilde{I}_n = \int dq \exp(-k_0^3\alpha q^2 x + iqk) = \sqrt{\frac{\pi}{k_0^3\alpha x}}\exp\left(-\frac{k^2}{4k_0^3\alpha x}\right). \tag{108}$$

All modes in this solution have equal intensities. The width of the angular spectrum has the following dependence on the propagation distance and the sound wave number:

$$\frac{\delta k}{k_0} = \sqrt{\alpha k_0 x}. \tag{109}$$

Equations (25) and (108) allow us to calculate the coherence function Γ_{nn} of the nth mode:

$$\Gamma_{nn}(x,\Delta y) = \frac{1}{b_n}\int dk e^{ik\Delta y}\tilde{I}_n(k) = \frac{2\pi}{b_n}\exp[-k_0^3\alpha x(\Delta y)^2] = \frac{2\pi}{b_n}\exp\left[-\left(\frac{\Delta y}{y_*}\right)^2\right], \tag{110}$$

where y_* is the horizontal coherence length, given by

$$y_* = \frac{1}{k_0\sqrt{k_0\alpha x}}. \tag{111}$$

IX. 2-D CASE

Let us consider the case of 2-D inhomogeneities, when

$$P_j(k_x,k_y) = P_j^{(1)}(k_x)\delta(k_y). \tag{112}$$

In this case $\sigma^{(2)}=0$ and Eq. (82) reduces to

$$\frac{\partial\tilde{I}_n}{\partial x} - i\frac{k\eta}{2b_n}\tilde{I}_n = -\beta_n\tilde{I}_n + \sum_m \sigma_{nm}^{(0)}\tilde{I}_m. \tag{113}$$

For $x\to\infty$, this equation has the following asymptotic solution:

$$\tilde{I}_n = I\exp\left(i\frac{k\eta}{2b_n}x\right), \tag{114}$$

where I is a constant. Substituting this expression into Eq. (25), one finds

$$\langle p_1 p_2^*\rangle|_{x-\mathrm{averaged}} = I\sum_n u_n(z_1)u_n(z_2)\int dk e^{ik\Delta y - iky}\int d\eta e^{-iky}I_n = I\sum_n u_n(z_1)u_n(z_2)\frac{1}{b_n}\int dk e^{ik\Delta y}\int d\eta e^{-i\eta y}\exp\left(i\frac{k\eta}{2b_n}x\right). \tag{115}$$

For $y=0$, we have

$$\langle p_1 p_2^*\rangle|_{x-\mathrm{averaged},y=0} = \frac{2I}{x}\sum_n u_n(z_1)u_n(z_2) = \frac{2I}{x}\delta(z_1 - z_2). \tag{116}$$

Introducing in Eq. (113)

$$\tilde{\tilde{I}}_n = \tilde{I}_n\exp\left(i\frac{k\eta}{2b_n}x\right), \tag{117}$$

we find

$$\frac{\partial\tilde{\tilde{I}}_n}{\partial x} = -\beta_n\tilde{\tilde{I}}_n + \sum_m \sigma_{nm}^{(0)}\tilde{\tilde{I}}_n. \tag{118}$$

Here,

$$\sigma_{nm}^{(0)} = \frac{\pi}{2}\left(\frac{\omega}{c_{00}}\right)^4\frac{1}{b_nb_m}\sum_j P_j^{(1)}(b_n - b_m)[N_{nm}^{(j)}(b_n - b_m)]^2. \tag{119}$$

Equations (118) and (119) agree with the corresponding equations in Ref. 4.

X. NUMERICAL RESULTS

A. Parameters of the problem

Numerical solutions of equations derived in this paper depend on many parameters: frequency and depth of a source, a model for sound speed fluctuations, an unperturbed sound speed profile, ocean depth, etc. Furthermore, these numerical solutions depend on parameters which determine numerical implementation, for example, discretization of the integrals over k and η, discretization of the IW spectrum, discretization of the integrals describing mode-mode interactions, a number of acoustic modes, and a value of the range step Δx. Therefore, studies of the dependence of numerical

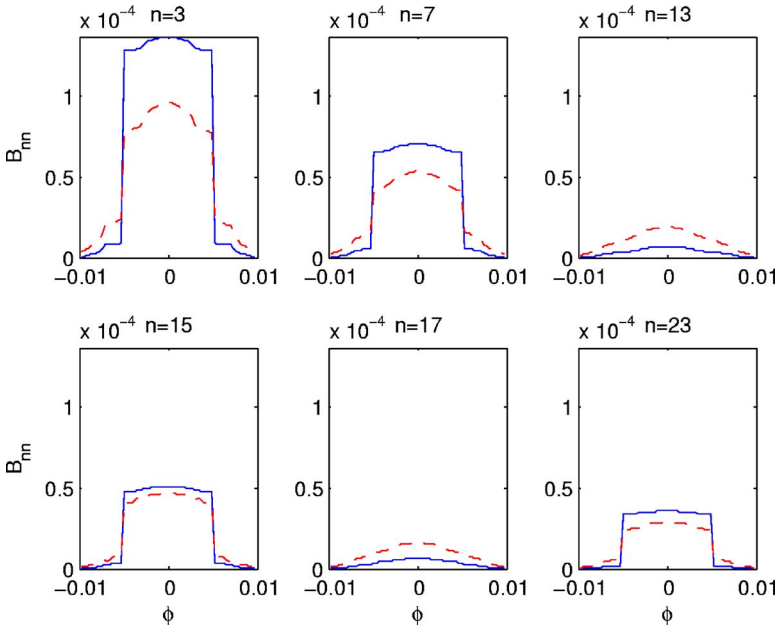


FIG. 2. (Color online) The dependencies of the mode correlation functions $B_{nn}(\phi,x,0)$ (in m) on the azimuthal angle ϕ (in rad) for modes $n=3,7,13,15,17,23$. Solid lines correspond to the propagation distance $x=250$ km, dashed lines to $x=1000$ km.

solutions on these parameters require extensive work and will be addressed in the future. In this section, we present two examples of numerical solutions of the derived equations: calculations of the coherence length and correlation functions.

When doing these calculations, we assumed that the Brunt-Väïsäla frequency $n(z)$ and the sound speed $c(z)$ are given by the Munk canonical profiles:

$$n(z) = n_0 \exp\left(-\frac{z}{B}\right), \tag{120}$$

$$c_0(z) = c_{00}\left\{1 + 0.0057\left[\exp\left(-2\frac{z-z_1}{B}\right) + 2\frac{z-z_1}{B} - 1\right]\right\}. \tag{121}$$

For numerical estimates, we used the following values:

$$n_0 = 5.2 \times 10^{-3}\mathrm{s}^{-1}, \quad c_{00} = 1500 \text{ m/s},$$

$$B = z_1 = 1000 \text{ m}.$$

We also assumed that the ocean depth is 3 km and that the spectrum of IWs is given by the Garrett-Munk spectrum:

$$P_j(q) = E_0 \frac{2}{\pi^3}\frac{j_* j}{j_*^2 + j^2}\frac{q_* q}{(q_*^2 j^2 + q^2)^2}, \quad q_* = \frac{\pi}{B}\frac{\omega_i}{n_0}, \tag{123}$$

with the following parameters:

$$E_0 = 4 \text{ m}^3/\text{s}^2, \quad j_* = 3, \quad \omega_i = 7.3 \times 10^5 \text{ s}^{-1}. \tag{124}$$

In Eq. (49), the IW modes $\Phi_j(q,z)$ were represented as follows:

$$\Phi_j(q,z) = \gamma_a \frac{n^2(z)}{n^2(z_1)} w_j(q,z), \tag{125}$$

where $\gamma_a = 1.14 \times 10^{-5}$ m^{-1} and $w_j(q,z)$ are the linear IW modes, satisfying the following equations:

$$\frac{d^2 w_j}{dz^2} + \left(\frac{n^2(z)}{\omega_j^2} - 1\right) q^2 w_j = 0,$$

$$\int (n^2(z) - \omega_i^2) w_j w_{j'} \, dz = \delta_{jj'}. \tag{126}$$

The smallest and the largest wavelengths of the IWs included into calculations of the interaction integrals were 100 m and 10 km, respectively.

Finally, it was assumed that the point source with frequency 75 Hz is located at $z_0 = 807$-m depth that corresponds to the 1998–1999 NPAL experiment.[21]

B. Coherence length

First, we calculate the transverse coherence length y_*. In these calculations, the number of acoustic modes was $N_{\mathrm{mod}} = 54$ (which is the total number of water-borne modes), and the number of the IW modes included into summation in Eq. (85) was $N_{\mathrm{IW}} = 100$. In Eqs. (95) and (96), integration over k was extended up to $\sin^{-1}(k/k_0) = 0.15$. As a result of these calculations, we obtained the following value of the parameter α:

$$\alpha = 4.94 \times 10^{-11}. \tag{127}$$

Let $x = 4 \times 10^6$ m, which corresponds to the sound propagation distance in the 1998–1999 NPAL experiment.[21] Using Eqs. (127) and (109), we can calculate the width of the angular spectrum of the acoustic field:

$$\frac{\delta k}{k_0} = 7.9 \times 10^{-3}. \tag{128}$$

Furthermore, using Eq. (111), we calculate the horizontal coherence length

$$y_* = 400 \text{ m}. \tag{129}$$

These values of α and y_* agree with those obtained by processing of the NPAL data.[21]

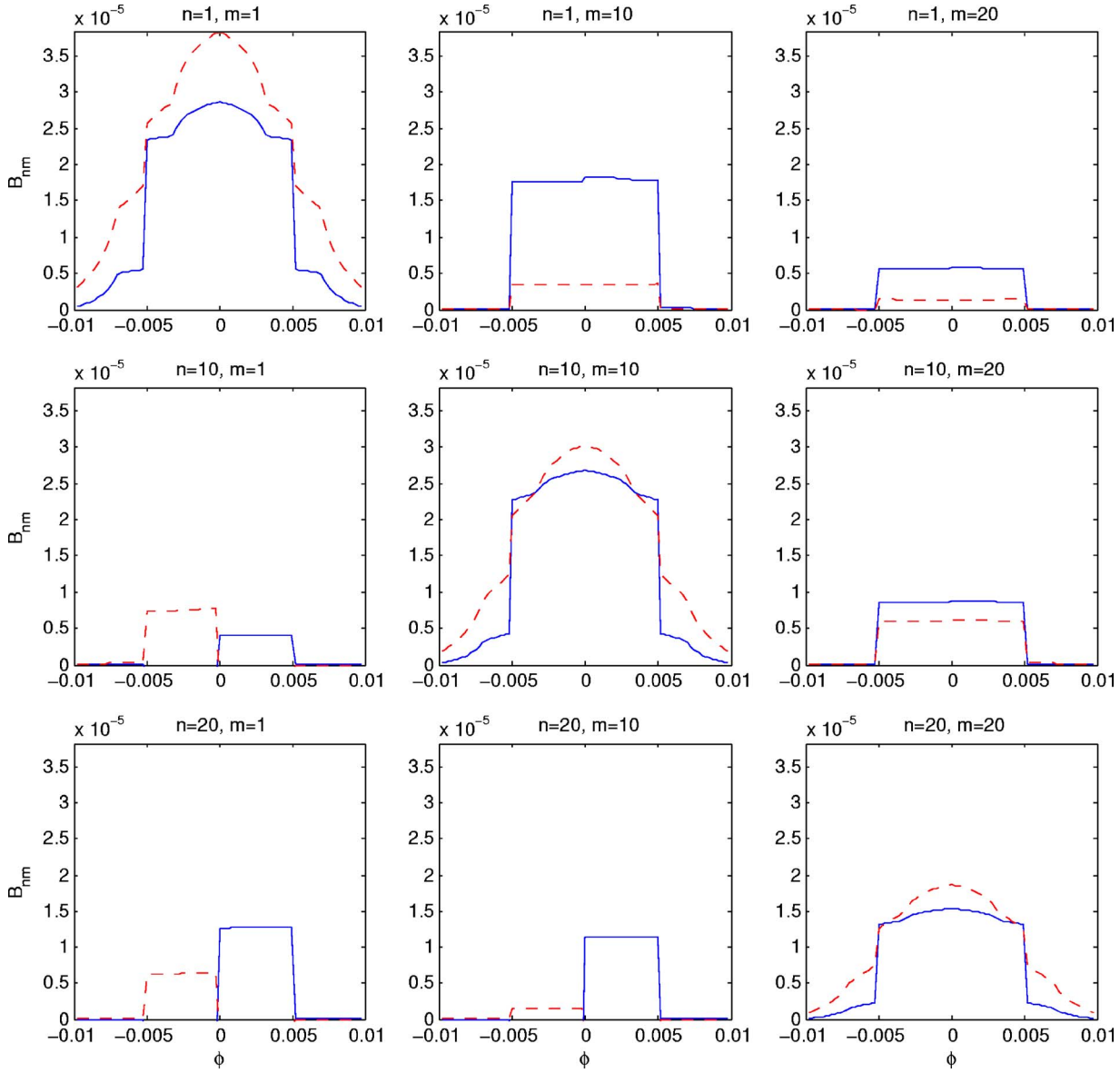


FIG. 3. (Color online) The dependencies of the cross-modal correlation functions $B_{nm}(\phi,x,0)$ (in m) on the azimuthal angle ϕ (in rad) for modes $n,m=1,10,20$. Solid lines correspond to the propagation distance $x=250$ km, dashed lines to $x=1000$ km.

C. Correlation functions

Here, using Eq. (74) we calculate the mode correlation functions B_{nn} and the cross-modal correlation functions B_{nm}. Since these numerical calculations are more intense than those in the previous subsection, we used 25 acoustic modes and 50 IW modes.

In Eq. (74), k was replaced with ϕk_0, where ϕ is the azimuthal angle. The integral in this equation was approximated as a sum over 100 components of ϕ within the range $(-0.01,0.01)$ rad. Furthermore, η was set to 0. [According to Eq. (31), this corresponds to the y-averaged coherence functions.] The length of a range step was $\Delta x=25$ km and 40 steps were used which covered 1000-km propagation distance. It was assumed that the source emits a narrow beam in the horizontal plane with uniform distribution within azimuthal angles $-0.005\leqslant\phi\leqslant0.005$ rad. In the absence of scattering this angular distribution would be kept constant for all propagation distances.

The numerical results obtained are presented in Figs. 2 and 3. Figure 2 shows the dependencies of the mode correlation functions $B_{nn}(\phi,x,0)$ on the azimuthal angle ϕ for modes $n=3,7,13,15,17,23$ and propagation distances $x=250$ km (solid line) and $x=1000$ km (dashed line). Note that at $x=0$ km the dependencies of B_{nn} on ϕ (not shown in Fig. 2) are rectangulars with the base $-0.005\leqslant\phi\leqslant0.005$ rad and different heights due to different excitation amplitudes of the modes by the source. It follows from Fig. 2 that the initial rectangular dependencies spread with distance at different rates for different modes. Furthermore, one can see that, at $x=1000$ km, the energy equipartization between modes is far from being reached yet.

Figure 3 shows the cross-modal correlation functions $B_{nm}(\phi,x,0)$ vs. ϕ for the same propagation distances and modes $n,m=1,10,20$. In the figure, the subplots above the main diagonal of the 3×3 matrix of pictures depict the real parts of the cross-modal correlation functions, $\mathrm{Re}\,B_{nm}$, and

those below correspond to the imaginary parts, Im B_{nm}. The diagonal subplots represent the mode correlation functions B_{nn} and are similar to those shown in Fig. 2. Note that the cross-modal correlation functions are smaller than the mode correlation functions. However, the cross-modal correlation functions are not extinct at the 1000-km propagation distance.

XI. CONCLUSIONS

In this paper, using the Chernov approximation, a theory of low-frequency, long-range sound propagation through a fluctuating ocean was developed. In particular, closed, finite-difference equations for the first two statistical moments of the sound field were derived. Simplified versions of these equations were obtained that account for the narrowness of the angular spectrum of the scattered acoustic field. One of the main results obtained is the equation for the correlation functions of the modes amplitudes, Eq. (74). Using this equation, mode correlations functions and cross-modal correlation functions were calculated numerically.

Spatial averaging of Eq. (74) allowed us to obtain an analytical solution of this equation. Using this approximate analytical solution, the horizontal coherence length of the acoustic field was estimated. The value of the horizontal coherence length was found to agree with that measured during the 1998–1999 NPAL experiment.

Finally, note that all approximate equations derived in the paper satisfy energy conservation.

ACKNOWLEDGMENTS

This work was supported by the ONR Grant No. N0001404IP20009. We would like to thank anonymous reviewers whose comments allowed us to improve this paper.

[1]S. M. Flatte, R. Dashen, W. H. Munk, K. M. Watson, and F. Zachariasen, *Sound Transmission Through a Fluctuating Ocean* (Cambridge U. P., Cambridge, 1979).

[2]R. Dashen, "Path-integral treatment of acoustical mutual coherence functions for rays in a sound channel," J. Acoust. Soc. Am. **77**, 1716–1722 (1985).

[3]O. A. Godin, V. U. Zavorotny, A. G. Voronovich, and V. V. Goncharov, "Refraction of sound in a horizontally-inhomogeneous, time-dependent ocean," IEEE J. Ocean. Eng., in press (2005).

[4]L. Dozier and F. Tappert, "Statistics of normal mode amplitudes in a random ocean," J. Acoust. Soc. Am. **63**, 353–365 (1978).

[5]C. Penland, "Acoustic normal mode propagation through a three-dimensional internal wave field," J. Acoust. Soc. Am. **78**, 1356–1365 (1985).

[6]W. E. Kohler and G. C. Papanicolau, "Wave propagation in a randomly inhomogeneous ocean," in *Wave Propagation and Underwater Acoustics*, edited by J. B. Keller and J. S. Papadakis, Lecture Notes in Physics, Vol. **70** (Springer-Verlag, Berlin, 1977), pp. 153–223.

[7]V. V. Artel'nyi, V. D. Kukushkin, and M. A. Raevskii, "Energy and correlation characteristics of low-frequency sound waves in underwater sound channels," Sov. Phys. Acoust. **32**, 369–372 (1986).

[8]A. G. Sazontov and V. A. Farfel', "Calculation of low-frequency sound attenuation in the ocean with connection with scattering by internal waves," Sov. Phys. Acoust. **32**, 307–310 (1986).

[9]V. I. Tatarskii, *The Effects of the Turbulent Atmosphere on Wave Propagation* (Keter, Jerusalem, 1971).

[10]A. Ishimaru, *Wave Propagation and Scattering in Random Media* (Academic, New York, 1978).

[11]S. M. Rytov, Yu. A. Kravtsov, and V. I. Tatarskii, *Principles of Statistical Radio Physics. Part 4, Wave Propagation through Random Media* (Springer, Berlin, 1989).

[12]E. Yu. Gorodetskaya, A. I. Malekhanov, A. G. Sazontov, and N. K. Vdovicheva, "Deep-water acoustic coherence at long-ranges: Theoretical prediction and effects on large-array signal processing," IEEE J. Ocean. Eng. **24**, 156–170 (1999).

[13]A. G. Sazontov, "A quasi-classical solution to the radiation transport equation in a scattering medium with regular refraction," Acoust. Phys. **42**, 487–494 (1996).

[14]N. K. Vdovicheva, E. Yu. Gorodetskaya, A. I. Malekhanov, and A. G. Sazontov, "Gain of a vertical antenna array in a randomly inhomogeneous oceanic waveguide," Acoust. Phys. **43**, 669–675 (1997).

[15]E. Yu. Gorodetskaya, A. I. Malekhanov, A. G. Sazontov, and V. A. Farfel', "Effects of long-range sound propagation in a random inhomogeneous ocean on the gain loss of a horizontal antenna array," Acoust. Phys. **42**, 543–549 (1996).

[16]L. A. Chernov, *Waves in Randomly-Inhomogeneous Media* (Nauka, Moscow, 1975) (in Russian).

[17]J. L. Codona, D. B. Creamer, S. M. Flatte, R. G. Frehlich, and F. Henyey, "Moment-equation and path-integral techniques fow wave propagation in random media," J. Math. Phys. **27**, 171–177 (1986).

[18]A. G. Voronovich, "Low-frequency sound propagation through random internal waves with application to measurements of internal wave spectra by acoustic means," in *Proceedings of the Fourth European Conference on Underwater Acoustic*, edited by A. Alippi and G. B. Canelli, Rome, Italy, 1998, Vol. **2**, pp. 751–756.

[19]S. Frankenthal and M. J. Beran, "Propagation in random stratified waveguides—A modal-spectral treatment," J. Acoust. Soc. Am. **104**, 3282–3295 (1998).

[20]J. A. Colosi, "Coupled mode theory for sound propagation through randominternal wave fields," J. Acoust. Soc. Am. **116**, 2535 (2005).

[21]A. G. Voronovich V. E. Ostashev, and NPAL Group, "Horizontal refraction of acoustic signals retrieved from the North Pacific Acoustic Laboratory billboard array data," J. Acoust. Soc. Am. **117**, 1527–1537 (2004).

[22]A. Voronovich, *Wave Scattering from Rough Surfaces*, 2nd ed. (Springer-Verlag, New York, 1999).

An experimental demonstration of blind ocean acoustic tomography[a)]

Sérgio M. Jesus[b)] and Cristiano Soares
SiPLAB - FCT, Universidade do Algarve, Campus de Gambelas, PT-8005-139 Faro, Portugal

Emanuel Coelho
NATO Undersea Research Centre, Viale San Bartolomeo 400, I-19138 La Spezia, Italy

Paola Picco
ENEA, Marine Environment Research Centre, P.O. Box 224, I-19100 La Spezia, Italy

(Received 23 June 2005; revised 13 December 2005; accepted 14 December 2005)

Despite the advantages clearly demonstrated by ocean acoustic tomography (OAT) when compared to other ocean monitoring techniques, it suffers from several technical-related drawbacks. One is the requirement for rather expensive equipment to be maintained and operated at several locations in order to obtain sufficient source–receiver propagation paths to cover a given ocean volume. This paper presents the preliminary feasibility tests of a concept that uses ships of opportunity as sound sources for OAT. The approach adopted in this paper views the tomographic problem as a global inversion that includes determining both the emitted signal and the environmental parameters, which is a similar problem to that seen in blind channel identification and was therefore termed blind ocean acoustic tomography (BOAT). BOAT was tested on a data set acquired in October 2000 in a shallow-water area off the west coast of Portugal, including both active and passive (ship noise) data. Successful results show that BOAT is able to estimate detailed water column temperature profiles coherent with independent measurements in intervals where the uncontrolled source signal (ship noise) presents a sufficient bandwidth and signal-to-noise ratio, which clearly define the limitations of the presented method. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2165070]

PACS number(s): 43.30.Wi, 43.30.Pc, 43.60.Cg, 43.60.Pt [AIT] Pages: 1420–1431

I. INTRODUCTION

Ocean acoustic tomography (OAT) was proposed more than three decades ago but its routine usage is still not a reality today. Despite the advantages clearly demonstrated by OAT when compared to other ocean monitoring techniques, it suffers from several drawbacks, such as the requirement for rather expensive equipment to be maintained and operated simultaneously at several locations in order to obtain sufficient source–receiver propagation paths to cover a given ocean volume. In particular, OAT is known to be very sensitive to the precise knowledge of source–receiver relative locations at all times. Passive acoustic tomography (PAT) is an acoustic tomography variant where the usual cooperative acoustic source is replaced by a noncooperative noise source as for example a ship of opportunity. The basic idea behind PAT is to extend the application of acoustic tomography to areas with heavy or regular ship traffic and where it would be impossible, or too costly, to deploy a controlled acoustic source in a permanent basis. PAT is also an interesting alternative to active tomography in the presence of marine mammals or for covert military application. The broad concept of using alternative illuminating sources for OAT or for geoacoustic inversion is not new. It was first proposed by Buckingham to use acoustic daylight to form images of silent objects in the ocean[1,2] and then using ambient noise for geoacoustic inversion.[3] More recently, Harrison[4–6] used sea surface wind-induced noise; then, Buckingham *et al.*[7] used light aircraft air-induced noise, both with the purpose of shallow-water geoacoustic inversion.

PAT differs from classic active acoustic tomography by the fact that in PAT, the source signal is stochastic with unknown characteristics and uncontrolled by the experimenter. There are at least two important implications of the assumptions made under PAT: one is that the emitted signal is possibly fluctuating over time both in strength and bandwidth, the other is that the sound source's position is (in general) unknown and possibly changing over time. The fact that the source position is unknown implies that, apart from the sound-speed profile to be inverted for, the other propagation channel characteristics (e.g., bottom properties, water depth, etc.) are also unknown and have to be estimated together with the source position. An inverse problem where both the input signal and the channel are unknown is termed a blind deconvolution problem, and is common in the fields of wireless communications, geophysics, and in all problems where channel identification is required and where the input signal is not known (see Cadzow[8] for an overview). The generally adopted methodology is to use higher-order statistics and (in wireless communications) the cyclostationarity properties of the received signal.[9,10] Such methods have also been used in

[a)]Portions of this work were presented at the European Conference on Underwater Acoustics in July 2002, Gdansk, Poland.
[b)]Electronic mail: sjesus@ualg.pt

 0001-4966/2006/119(3)/1420/12/$22.50

underwater acoustics for signals with some degree of nonstationarity.[11,12] Assuming that the noise sources of opportunity are relatively stationary inputs to the propagation channel, it is possible to build a model-based cost function where both the source and the channel properties are unknown variables to be estimated. In essence, due to the highly random and incoherent nature of the signal phase along the propagation path, the actual emitted waveform is almost never used for matched-field processing at the receiver. The exception to this is of course coherent underwater communications. It is therefore sufficient to consider a second-order cost function working in a frequency band where the signal is assumed to propagate. Using this analogy between blind deconvolution in passive ocean tomography, this technique was termed blind ocean acoustic tomography (BOAT). The distinction between PAT and BOAT is that the former aims at estimating ocean temperature with alternative passive sources, while the latter produces a full environmental estimate, including water column, bottom properties, and source–receiver geometry as well as a source-emitted power spectrum, without any knowledge or control on the acoustic illuminating source. Therefore, BOAT deals both with the emitting source power spectrum and with the estimation of a full environment model response, often described by a variable but large number of physical parameters.

The hardest problem in BOAT is dealing with a large number of parameters. Difficulties are associated not only with the number of parameters but also with the different nature of the parameters in play, where geometrical, water column, and bottom parameters are simultaneously searched for. An additional difficulty arises from the fact that our degree of knowledge of the parameters to be inverted for may be highly variable. Several approaches have been proposed to deal with such inverse problems either for geoacoustics,[13–16] or for the water column.[17,18] In other studies the parameters under search simultaneously included geoacoustic, water column, and geometric parameters both known and unknown.[19,20] The term generally used for approaches performing multiparameter search including both known and unknown parameters is *focalization*, and was first proposed by Collins *et al.*[21] in the context of range-depth source localization. So, in our case the problem to be addressed under the scope of this paper involves an environmental and geometrical focalization procedure with a random source of unknown spectra. It is well known that conventional matched-field processing (MFP) Bartlett-based cost functions are suboptimal when dealing with highly variable (nonflat frequency spectrum) power sources.[22] An alternative procedure for estimating also the source-emitted spectra using the estimates as weighting function for the conventional MFP-based cost function was proposed by the authors in Ref. 23 and is used here in the INTIFANTE'00 data set.

The INTIFANTE'00 sea trial was a joint experiment carried out by Instituto Hidrográfico and the University of Algarve and the collaboration and support of several other institutions [Insituto Superior Técnico (IST), Lisboa, Portugal, Ente per l'Energia ed l'Ambiente (ENEA) and SACLANT Undersea Research Centre, both in La Spezia, Italy] in the fall of 2000, in a shallow-water area 50 km south from Lisbon, Portugal. There was a broad range of objectives pursued by this experiment, among which the testing of the ability of inverting ocean properties with both known and unknown, active and passive, stationary and moving acoustic source signals in various environments, both range independent and range dependent. The results shown in this paper cover the following situations: range independent and range dependent with a moving unknown active source, and a fast-moving ship acting as source in a partially range-dependent environment.

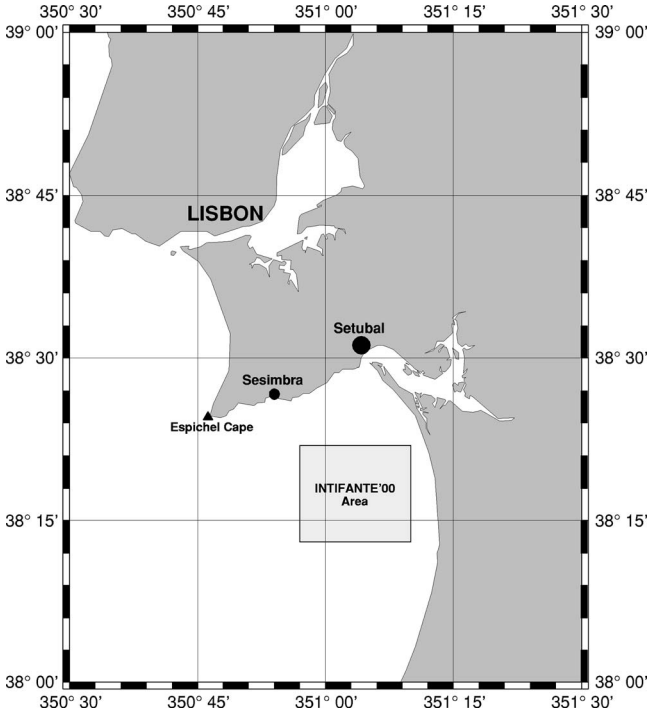


FIG. 1. INTIFANTE'00 sea trial location.

This paper is organized as follows: Sec. II gives an overview of the INTIFANTE'00 experiment. Section III presents the environmental model drawn from the direct measurements performed during the sea trial. Methods and algorithms used during the processing of the data are described in Sec. IV. Section V describes the results obtained with a towed active source and a ship, the research ship NRP D. Carlos I, both assumed unknown. Final conclusions and perspectives are drawn in Sec. VI.

II. THE INTIFANTE'00 SEA TRIAL

The INTIFANTE'00 (INTIFANTE is a made-up acronym from INTImate and inFANTE, two cofunding projects) sea trial was carried out in the vicinity of Setúbal, situated approximately 50 km to the south of Lisbon, in Portugal, during the period from 9 to 29 October, 2000 (Fig. 1). A detailed and complete description of the experiment and of the various data sets acquired during the INTIFANTE'00 sea trial can be found in Ref. 24, while here only a brief description will be given. The experiment area was a rectangular box situated within the continental shelf with depths varying from 60 to 130 m. The actual acoustic runs were performed

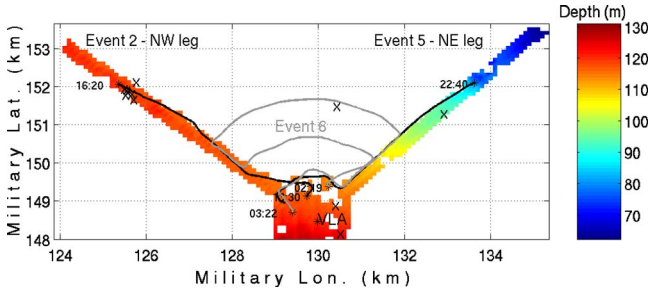


FIG. 2. (Color online) INTIFANTE'00 sea trial: acoustic runs and bathymetry during events 2, 5, and 6. X signs mark the XBT locations and VLA indicates the vertical line array location.

over three distinct paths on whose intersection a vertical line array (VLA) was moored, as shown in Fig. 2. This paper reports the results obtained along the northwest (NW) path (event 2), along the northeast (NE) path (event 5) and between the NW and NE paths (event 6) (gray line in the figure). The path directed to the NW, parallel to the continental platform, is approximately range independent, while the NE path, oriented towards the coastline, is range dependent with water depths varying from 120 m at the VLA location to 70 m at the path end. In between these two paths, the environment is progressively and slowly changing from range independent to range dependent when going from NW to NE. As an overview of the technical aspects involved in the experiment, it can be referred that acoustic signals were transmitted with an acoustic transducer suspended from the research vessel NRP D. Carlos I, a Portuguese Navy oceanographic research ship managed by Instituto Hidrográfico (IH), and received on a 16-hydrophone 4-m-spacing VLA. The acoustic aperture of the VLA was located between the nominal depths of 30 and 90 m in a 120-m-depth water column. The acoustic signals received in the VLA were transmitted via an RF link to onboard ship, processed, monitored, and stored. The acoustic portion of the VLA was hanging from the sea surface and attached through a 70-m-long umbilical to the radio buoy that was itself bottom moored. Therefore, the VLA was navigating within a radius of approximately 70 m around the surface buoy mooring and moving up and down with surface motion and tides. This will be shown to be an important experimental detail during the analysis and validation of the acoustic data.

A. Source–receiver geometry and events

Figure 2 also shows the experimental site bathymetry along the acoustic paths together with the source ship tracks during these two events. At the beginning of event 2, the source ship started close to the VLA and then went off at low speed up to a range of approximately 5.5 km to the NW, transmitting precoded linear frequency modulated (LFM) signals. During event 5 the source ship approached the VLA from the NE along the range-dependent path while transmitting a pseudorandom noise (PRN) sequence. During event 6 the acoustic source was recovered and the ship itself was used as noise source. As seen on Fig. 2, during event 6, the ship started its run close to the VLA, went off to the NE leg for about 3.2 km, and then performed a sharp turn to the

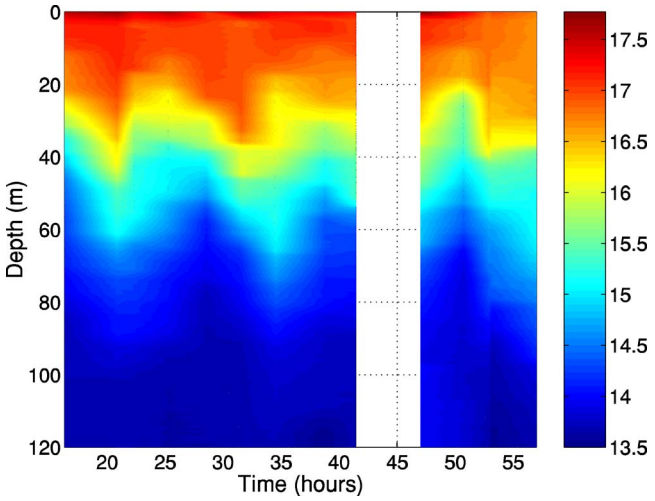


FIG. 3. (Color online) XBT cast temperature profiles.

west along a circular pattern around the VLA (gray line). The ship then repeated this pattern twice at VLA ranges of 2.2 and 1.2 km, closing to the VLA. During event 6 the research vessel NRP D. Carlos I was steaming at her maximum speed of 10 kn, in an attempt to generate as much noise as possible.

B. Environmental data measurements

Environmental data include direct measurements performed during the cruise, like water column temperature and bathymetry along runs, as well as archival data with geological information of the area. Water column temperature was gathered from XBTs and a thermistor chain collocated with the VLA. XBTs were sparse in time, approximately one every 3 h in order to capture tidal evolution, and spatially distributed throughout the experiment site. See Fig. 2 for XBT location and Fig. 3 for the recorded temperature profiles. Thermistor chain recordings provide a high time resolution but a few observation samples in depth and at a single spatial location. The VLA has 8 temperature sensors at 8-m spacing. The first sensor is located 3 m below the shallowest pressure gauge that was recording a mean depth of 30 m. So, temperature was recorded at approximate depths of 33, 41, 49, 57, 65, 73, 81, and 89 m, depending on the precision of the pressure gauge sensor and tilt of the VLA. The temperature field is shown in Fig. 4(a)–4(c) for events 2, 5, and 6, respectively. All three events are of relatively short durations (between 45 min and 1 h 30 min), which makes temperature recordings almost featureless. In all cases temperature is varying between 14 and 16.5 degrees Celsius with, in some cases, a few thermocline oscillations at the minimum observed depth of 30–35 m. Whether those temperature oscillations are due to effective thermocline changes or to VLA vertical movements is unknown.

The acoustic transmission tracks were surveyed during a previous sea trial in 1999 with both a sidescan sonar and a light seismic Sparker system. During INTIFANTE'00 a second sidescan sonar survey was performed. The overall conclusion is that the NW range-independent track used during event 2 has a mean depth of 119 m with a variation of ±3 m after tide correction, while the bottom is composed of fine

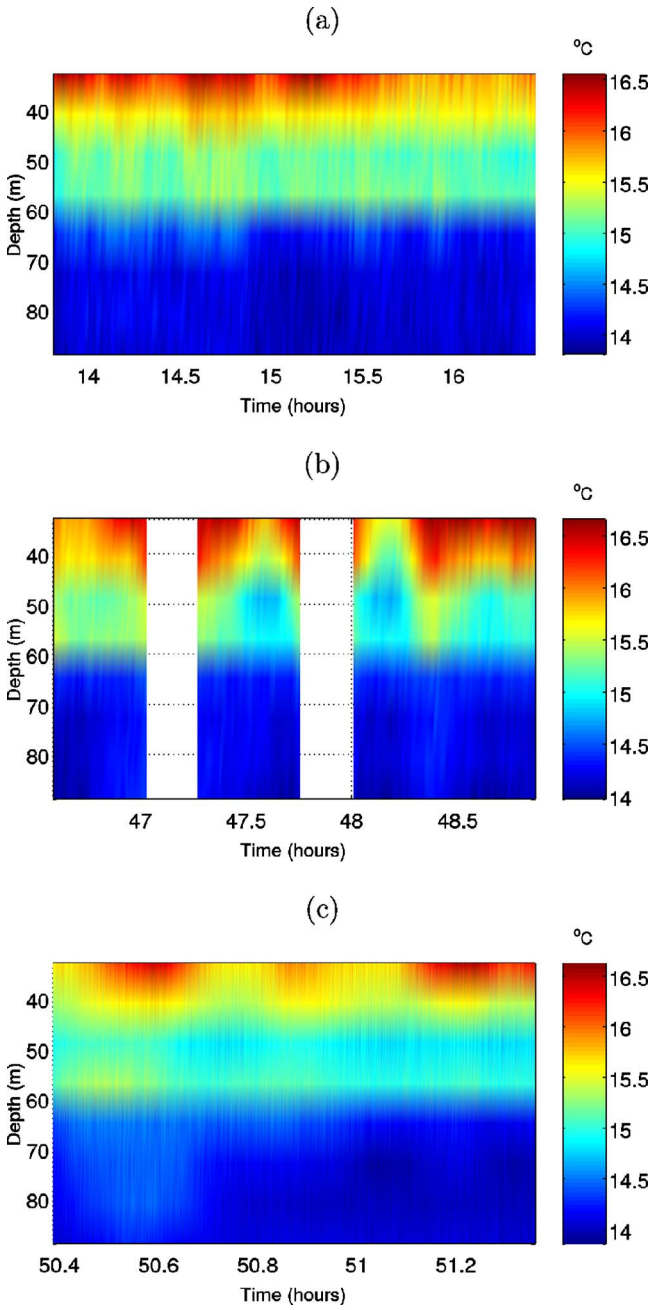


FIG. 4. (Color online) Temperature field recorded at the VLA during event 2(a), event 5(b), and event 6(c).

sand with a sediment layer of variable thickness between 4 and 6 m, in the first 6 km from the VLA. The NE leg has highly variable bottom properties with mud and hard rock patches.

C. Experiment geometry

Source–receiver range is calculated from the known VLA mooring position and the ship GPS log and is superimposed with the estimated source–range results (see the figures of Sec. V). An additional source of information is the recording of ship speed and bearing during event 6, as shown in Fig. 5, where it can be seen that the ship (acting as noise source in this event) maintained a mean speed of approximately 9 kn with several speed drops during sharp turns.

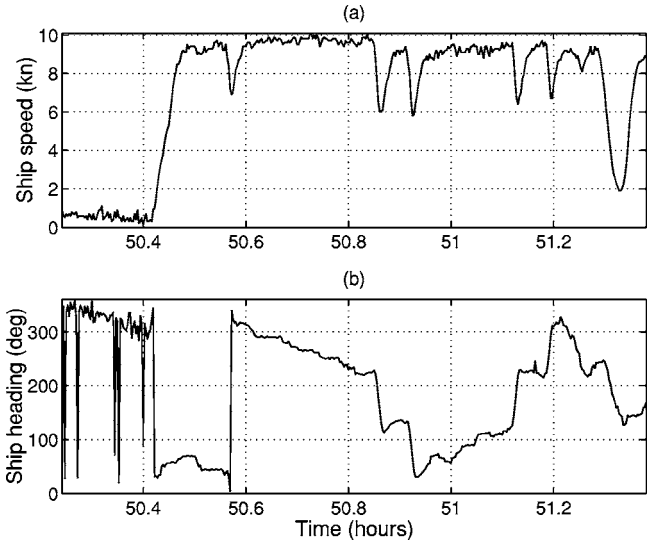


FIG. 5. GPS estimated ship speed (a) and ship heading (b) during event 6.

Unfortunately, the depth sensor collocated with the sound source failed to work during the whole experiment, so only information from cable scope and source depth variation versus tow cable length for various tow speeds was available to estimate source depth during the various runs.

III. ENVIRONMENTAL MODELING

An important first step in tomographic inversion is the choice of an environmental model able to represent the mean characteristics of the media where the signal is propagating. Such model will be called the baseline model and generally includes the available *a priori* information for the problem at hand. Besides the column temperature measurements and the sidescan sonar survey described in the previous section, there were no other *in situ* geological observations (e.g., cores) that could be used as additional background information, specifically for setting up the model bottom characteristics. Thus, it was decided to adopt a very generic baseline environmental model consisting of an ocean layer overlying a sediment and a bottom half-space assumed to be range independent, as shown in Fig. 6. The geoacoustic properties were drawn from generic geological knowledge of the area where it was assumed that the NW range-independent track had a quite regular bottom, covered by fine sand. Since event 5 is made along the NE track, a range-dependent version of the environmental model shown in Fig. 6 was used, where the bathymetry is simulated by a uniform bottom slope with a water depth varying from 70 m at the source location to 119 m at the VLA location. Event 6 takes place between the NW and NE tracks, where the latter is range dependent. In this case a range-independent model was used and therefore some degree of environmental model mismatch is anticipated.

Another important problem when inverting acoustic data is the difficulty associated with the representation of the sound-speed field in time, depth, and range by a finite set of invariant parameters. The solution for this problem is known as data regularization, and it consists of the expansion of the temperature, or equivalently, the sound-speed field, if the sa-

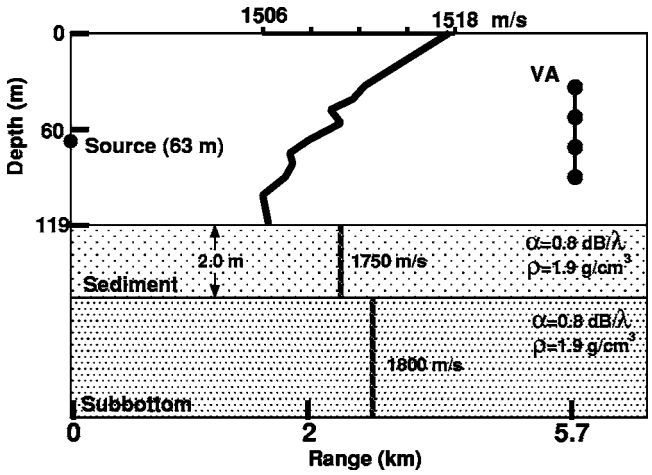


FIG. 6. Baseline environmental model for the range-independent propagation scenario of event 2.

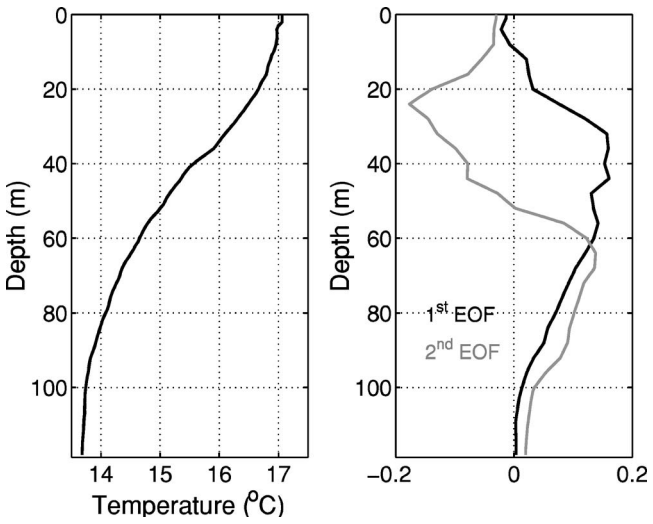


FIG. 7. XBT-based data used for temperature estimation: mean temperature profile (left) and first two empirical orthonormal functions (EOFs) (right).

linity profile in the area is known or can be represented with some degree of accuracy by a constant value over depth, on a basis of functions representative of the data set to be estimated. Since the inverse problem is constrained by the basis functions used, it is of paramount importance to determine a suitable function basis for each data set. It is well known that regularization is a powerful method that leads to a faster convergence and a higher uniqueness of the optimal solution, since prior information already contained in the set of basis functions allows one to start the search closer to the optimal solution than if no such information was available. A well-known method for obtaining a suitable basis function is to calculate the empirical orthogonal functions (EOF) as the eigenfunctions of the data correlation matrix. Assuming that the temperature field is stationary in range, this method requires a time series representative of the data under observation taken in the location and at the time of the experiment. In other words, in order to obtain a tomographic image of the temperature data one needs to already have an observation of those data. In many situations, and also in passive tomography, that is a strong *a priori* requirement that is not always fulfilled. Alternatively, it has been shown that a suitable set of eigenfunctions can be formed from the hydrodynamical normal modes (HNM) of oscillation of the temperature field.[25] In that case only a mean temperature profile from archival data is necessary.

In our case, since a full set of observations was readily available, the EOF method was used for parametrization of the ocean temperature. The EOFs were obtained using a singular value decomposition (SVD) of a data matrix **C** with columns

$$\underline{C}_i = \underline{T}_i - \underline{\bar{T}}, \tag{1}$$

where $\underline{T}_i$ are the real profiles available, and $\underline{\bar{T}}$ is the average profile. The SVD is known to be given by

$$\mathbf{C} = \mathbf{UDV}, \tag{2}$$

where **D** is a diagonal matrix with the singular values, and **U** is a matrix with orthogonal columns, which are used as the EOFs. The temperature profile is obtained by

$$\underline{T}_{\mathrm{EOF}} = \underline{\bar{T}} + \sum_{n=1}^{N} \alpha_n \underline{U}_n, \tag{3}$$

where N is the number of EOFs to be combined, judged to accurately represent the temperature field for the problem in hand. Generally, a criterion based on the total energy contained on the first N EOFs is used. Experimental results have shown that usually the first 2 or 3 EOFs are enough to achieve a high degree of accuracy. The 14 temperature profiles obtained from the XBT measurements (see Ref. 24) served as a database for the computation of the EOFs.

The criterion used to select the number of relevant EOFs for the available data was

$$\hat{N} = \min_{N} \left\{ \frac{\sum_{n=1}^{N} \lambda_n^2}{\sum_{m=1}^{M} \lambda_m^2} > 0.8 \right\}, \tag{4}$$

where the λ_n are the singular values obtained by the SVD, M is the total number of singular values, provided that $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_M$. For this data set criteria (4) yielded N=2, i.e., the first two EOFs are judged sufficient to model the sound speed with enough accuracy (see Fig. 7). The coefficients α_n, which are the coefficients of the linear combination of EOFs, are now part of the search space, i.e., they are searched as free parameters.

IV. FOCALIZATION: A METHOD FOR GLOBAL INVERSION

Multiple environmental and geometrical parameter optimization is often a computationally cumbersome task. The optimization hypersurface may be very irregular, leading to a severe ill-conditioned problem with a large number of local extrema. When dealing with real data, the inherent model mismatch and the presence of noise create a situation where there is no assurance of existence of an optimum solution in coincidence with (or even close) to the true model parameters. The first approach to the problem is to try to get as much *a priori* information as possible from the environmen-

tal parameters into the baseline model, in order to set them fixed and close to the true parameters, so the search is only done on a few unknown parameters. In practice, it is well known that setting fixed parameters in the model creates severe mismatches with real data that cannot be overcome by the search parameters, leading to poor fit situations and strongly biased estimates. An alternative is a technique proposed by Collins *et al.*,[21] known as focalization, where a number of *a priori* known model parameters, are allowed to be adjusted during the search process in order to compensate for the data-model misfit and possible measurement errors. Generally, *a priori* known parameters have a smaller degree of variation than the unknown parameters. This technique provides a high degree of adaptivity, a higher model fit, and a better-conditioned convergence to the true parameter values. Examples are given in recent publications by the authors.[18,20,26,27]

The optimization technique for reducing the number of forward computations was based on a genetic algorithm (GA). Principles of GA are now well known in the underwater acoustic community and elsewhere, and various strategies have been widely used in practice with positive results. GA parameter setting may have a strong influence on convergence to the solution and hence on the final result. These parameters have been set differently for each case and are explained below. The GA implementation used in this study was proposed in Ref. 28. In particular, a new scheme that was found to drastically optimize the search is to use the final solution at a given time period in the initialization of the solving procedure of the next time period.

Let us assume that at time t_i the best individual of the last population is $b(t_i)$. The GA is initialized at time t_{i+1} such that 30% of the individuals of the initial population are uniformly distributed within a 10% variation interval of the coordinates of $b(t_i)$. The other 70% are randomly distributed in the whole search space, in order to maintain a high degree of diversity. With this procedure the number of iterations has been decreased at each time period except for the first one. In practice it is verified that the model fit drops at the beginning of each time period when compared with its value at the end of the previous time period, which denotes that the data have changed and that the model is catching up through the adjustment of the parameters. However, after that initial fitness drop, rapid convergence to their "right" values, or at least to those giving the highest fit, is obtained. The objective function used in this study was based on the incoherent Bartlett processor in a frequency band selected according to the sound source response function. In order to be more specific, let us recall the Bartlett power function

$$B(\theta,t_n)=\frac{1}{K}\sum_{k=1}^{K}\mathbf{w}(\omega_k,\theta)^H\hat{\mathbf{R}}_Y(\omega_k,t_n)\mathbf{w}(\omega_k,\theta), \tag{5}$$

where $\hat{\mathbf{R}}_Y(\omega_k,n)$ is the data cross-correlation matrix estimate at frequency ω_k and at time interval t_n, $\mathbf{w}(\omega_k,\theta)$ is the model replica acoustic pressure at frequency ω_k and for search parameter θ, K is the number of frequencies, and ^{H} indicates conjugate transpose. The number and values of the selected frequencies varies with each case and as a general rule it was a compromise between the degree of difficulty of the problem at hand and the computational load. The C-SNAP code[29] was used as forward model.

V. RESULTS

In order to gain some insight into the processing of the ship noise data for the purpose of BOAT, it was decided to first test methods and algorithms with the active source data along both the NW range-independent and the NE range-dependent tracks, respectively, described in Secs. V A and V B. Since the ship noise data was obtained with a (fast) moving ship it was decided to analyze parts of data of these two events involving moving sources.

A. Active data in a range-independent track: Event 2

During event 2, a series of acoustic 170–600-Hz LFM sweeps was transmitted over a range-independent shallow-water waveguide, while the source was towed away from the VLA location. Results are shown in Fig. 8, and the following comments apply: the Bartlett power [given by (5)] is relatively high throughout the run, source range is well estimated, and all the other parameters are jointly estimated with credible values, including bottom properties. The temperature profiles are modeled by two EOFs whose coefficients show a smooth evolution through time, giving rise to a nicely stratified temperature estimate. Note that the environmental cross section being inverted for changes with time since the source is moving away from the VLA which, to some extent, may explain the observed high variability of non-time-varying parameters such as sediment and bottom sound speed as well as sediment thickness.

B. Active data in a range-dependent track: Event 5

During event 5 the sound source was emitting a pseudo-random noise (PRN) sequence in the band 150–1100 Hz, supposed unknown at the receiver. The inversion results are shown in Fig. 9. This run is a good example on how the three indicators—source range, source depth, and Bartlett power—can be used to validate environmental model estimates. At the beginning of the run, until time 47.04, the Bartlett power varies between 0.4 and 0.8, source range changes rapidly, most of the other parameters have highly variable values, and some are on, or near, the bounds of their search intervals. So, in this initial period, temperature estimates [plot (k)] cannot be considered as valid. At time 47.04, source range is suddenly estimated very close to the true values at 4-km range and steadily follows the approaching of the source to the VLA up to time 48 at about 2-km source range. During that interval most of the parameters, except the sediment thickness and the EOF coefficient α_1, follow stable values well within their respective intervals and are therefore mostly credible. The first EOF coefficient suffers a strong, and to date unexplained, change at time 47.52 right in the middle of that smooth path. After time 48, when the source has reached the closest point of approach to the VLA, the model match is again suddenly lost with strong variations on all parameters: drop of the Bartlett power from 0.8 to 0.3, a sudden range

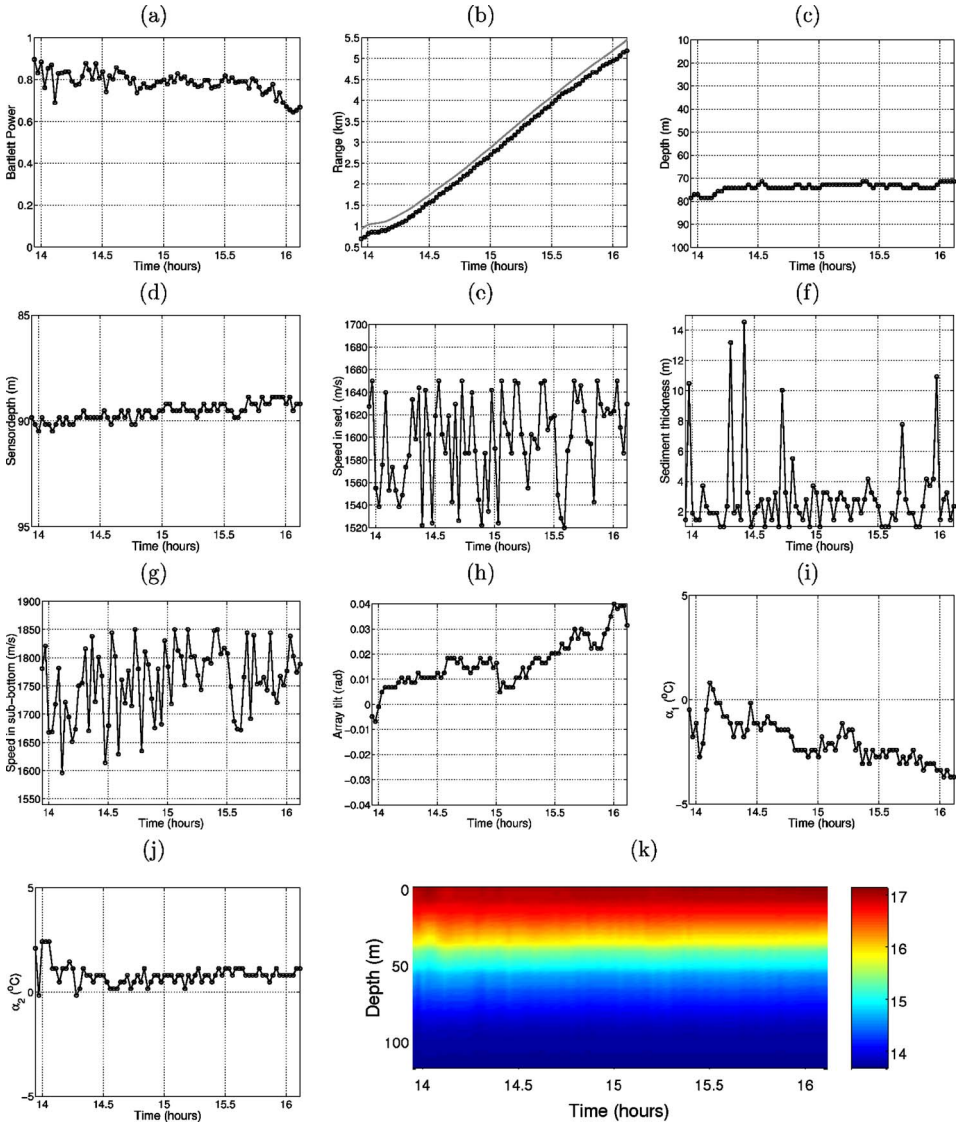


FIG. 8. (Color online) Focalization results for event 2: Bartlett power (a), source range (b), source depth (c), receiver depth (d), sediment compressional speed (e), sediment thickness (f), sub-bottom compressional speed (g), VLA tilt (h), EOF coefficient α_1 (i), EOF coefficient α_2 (j), and reconstructed temperature profiles (k).

variation from 1 to 3.5 km, and a drop of 10 m on source depth. After the work of Collins,[21] and using an analogy with optics, the terms "focus" and "out of focus" are often used to designate these alternate model adjustments and misadjustments. It was found that time 48 coincides with the low-tide change producing a 1.5-m rise on the array accompanied by strong variations of array tilt, as measured on the depth sensors and tiltmeters on the VLA [see Fig. 4.9 plots (a), (b), and (c) of Ref. 30]. There is the belief that during the tide change the array assumed a shape that was not taken into account by the baseline model. The model regains stability after 15 min with smooth parameter estimates and high Bartlett power values. Among all obtained values within validated intervals, source range and depth were clearly in agreement with the expected values; sound speed in the sediment and bottom are reasonably well estimated to have expected mean values of 1580 and 1700 m/s, respectively, with a higher uncertainty in the latter; and finally array depth and array tilt are in good agreement with the pressure and tilt sensors collocated with the VLA. After focalization the water temperature was reconstructed—plot (k)—showing a highly perturbed estimate due to successive focus and loss of focus through time. Also in this case the environment cross section is changing along time, possibly even more than in event 2, since the environment is range dependent with patches of rock and mud as reported in geological charts of the area; therefore, an even higher variability of bottom parameters is anticipated.

C. Passive data in a partially range-dependent track: Event 6

The real challenge comes when addressing the problem of tomographic inversion using ship noise data, i.e., a real unknown and stochastic source signal at unknown location and moving in a poorly known environment. This chapter addresses this problem using as example the data gathered

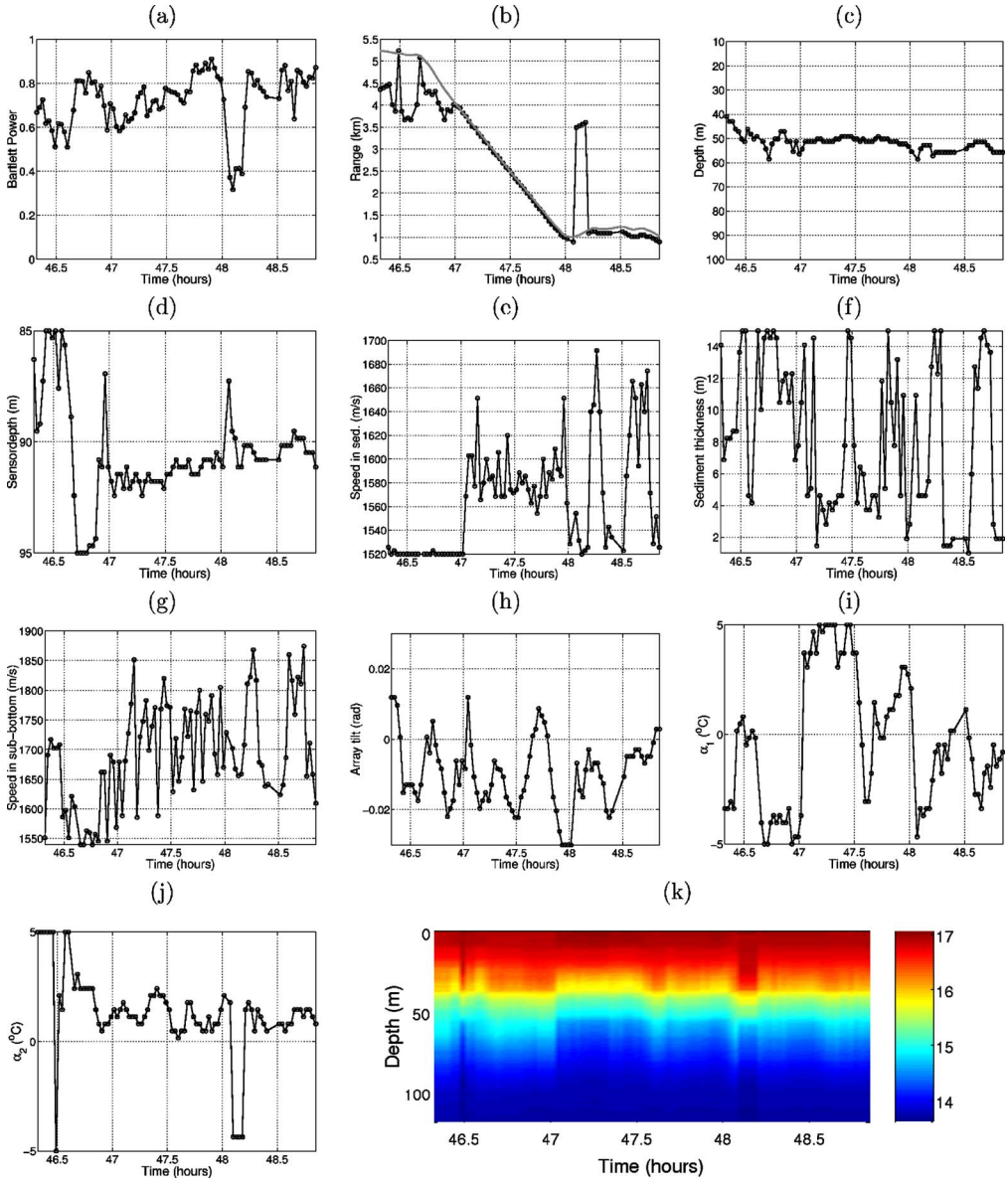


FIG. 9. (Color online) Focalization results for event 5: Bartlett power (a), source range (b), source depth (c), receiver depth (d), sediment compressional speed (e), sediment thickness (f), sub-bottom compressional speed (g), VLA tilt (h), EOF coefficient α_1 (i), EOF coefficient α_2 (j), and reconstructed temperature profiles (k).

during event 6, when the acoustic source was replaced by the NRP D. Carlos I herself as noise signal generator for tomographic inversion purposes.

1. Ship-radiated noise

Research vessel NRP D. Carlos I is a 2800-ton relatively recent ship, built in 1989, whose primary purpose was acoustic surveillance when it served under the U.S. flag. She has an overall length of 68 m and a beam of 13 m. Her main propulsion system is formed by two diesel-electric engines developing 800 HP attaining a maximum speed of 11 kn. According to her characteristics NRP D. Carlos I can be considered as an acoustically quiet ship. Hence, her use for the purpose of passive tomography can be considered as providing conservative results when compared with full-length cargo ships or tankers traveling at cruising speed. In order to maximize the probabilities of successful inversion and get close to the cruising speeds of "normal" ship traffic, NRP D. Carlos was set to steam at her full speed along a series of three concentric portions of circle as shown in Fig. 2 (gray line) with ship's speed and heading as shown in Fig. 5, plots (a) and (b), respectively. As an example, Fig. 10 shows a time-frequency plot of the relative power spectrum received on hydrophone 8 at 60-m depth (a), and a mean power spectrum over the whole event (b). There are clearly a few characteristic frequencies emerging from the background noise between 250 and 260 and a strong single tone at 359 Hz. There is also a colored noise spectra in the band 500 to 700 Hz with, however, a much lower power.

2. Environmental model

During this test several difficulties are added to the problem, when compared to the previously analyzed data set obtained in events 2 and 5: (i) the source is moving fast; (ii) the

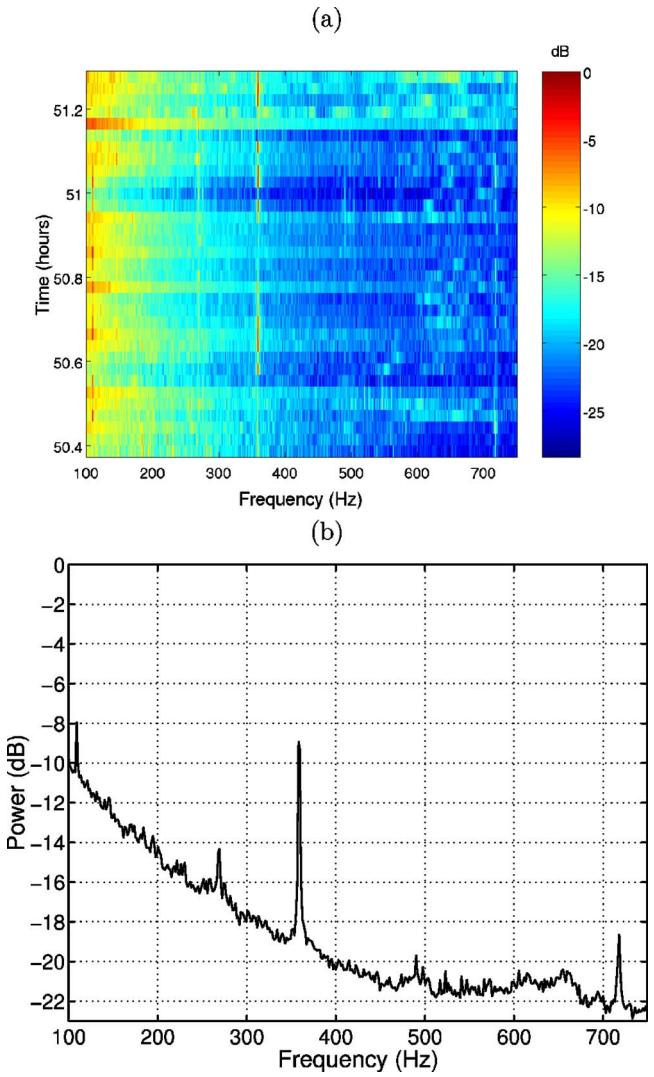


FIG. 10. (Color online) NRP D. Carlos I ship radiated noise received on hydrophone 8, relative power scale on a time-frequency plot (a) and mean power spectrum (b).

environment is a mixture of range dependent and range independent; and (iii) the source signal is ship noise with unknown and presumably time-varying characteristics. On top of those difficulties, the actual processing adds also a further problem which is that it is not possible to decide during the processing to switch between range-independent and range-dependent environmental models. In theory, a range-dependent model is also applicable to the range-independent case, allowing water depth at the source end to change along the ship track. Due to the well-known source range versus water depth inter-relation and parameter hierarchy, it is impossible, or at least extremely difficult, to simultaneously estimate source range and water depth as well as other low dependence parameters. In this analysis it was decided to use a range-independent model for reducing the computation and inversion burden. As will be seen in the following section, the usage of a range-independent model, even in a slightly range-dependent environment, will add a significant source range mismatch, at some well-defined points during the processing.

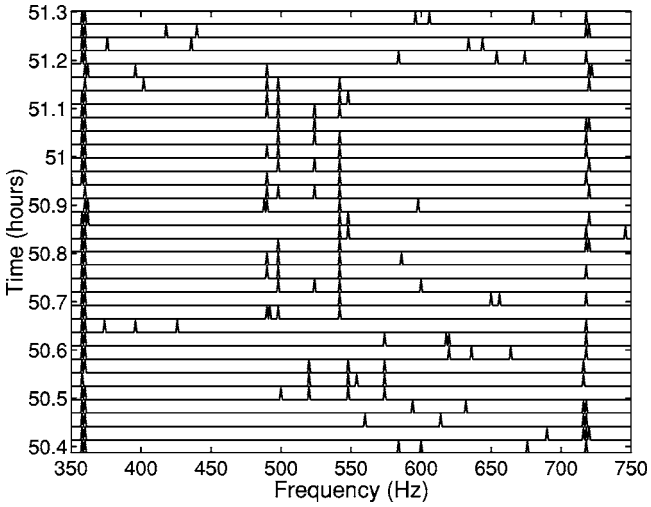


FIG. 11. INTIFANTE'00 sea trial, event 6 frequency selection: 1 - selection and 0 - no selection based on contiguous snapshot mean signal coherence along time [Eq. (6)].

3. Inversion results with a coherence-based frequency selection

The inversion methodology was based on a three-step procedure: (i) preliminary search of the outstanding frequencies in a given time slot; (ii) the usual parameter focalization, based on an incoherent broadband Bartlett processor; a C-SNAP (Ref. 29) forward acoustic model and a GA based optimization and (iii) inversion result validation based on model fitness and coherent source range and depth estimates through time. There are a number of possibilities for implementing the frequency selection procedure for the first step. A first attempt using a simple periodogram based spectral estimator for frequency selection was shown in Ref. 31 and a similar estimator coupled with an amplitude estimator for frequency selection and weighting in Ref. 23. In the present paper a coherence-based approach is proposed, where the time coherence of a given frequency ω is evaluated according to

$$\Gamma(\omega) = \frac{1}{N-1}\sum_{n=1}^{N-1} \frac{\mathbf{Y}^H(\omega,t_{n+1})\mathbf{Y}(\omega,t_n)}{\|\mathbf{Y}(\omega,t_{n+1})\|\|\mathbf{Y}(\omega,t_n)\|} \geqslant \gamma_0, \tag{6}$$

where $\mathbf{Y}(\omega,t_n)$ is an L-dimensional complex vector with the L array sensor output at frequency ω in time snapshot t_n, N is the total number of time snapshots in a given window, and γ_0 is a constant detection threshold depending on the actual SNR and signal/ambient noise level. The idea is that if a signal is present in a given frequency bin, a slow change of the channel structure would make this signal maintain its coherence from one snapshot to the other while there is a good chance that ambient noise will have a lower time coherence. For the data of event 6, 16 s of data were divided into 0.5-s-duration snapshots, thus giving a frequency resolution of 2 Hz, and a number of snapshots $N=32$. The result of applying (6) in the frequency band 350 to 750 Hz gave the results shown in Fig. 11. This figure shows which bins were selected for processing—the third axis is 1 upon selection, and 0 otherwise. Note that there are a couple of frequencies almost constantly present through-

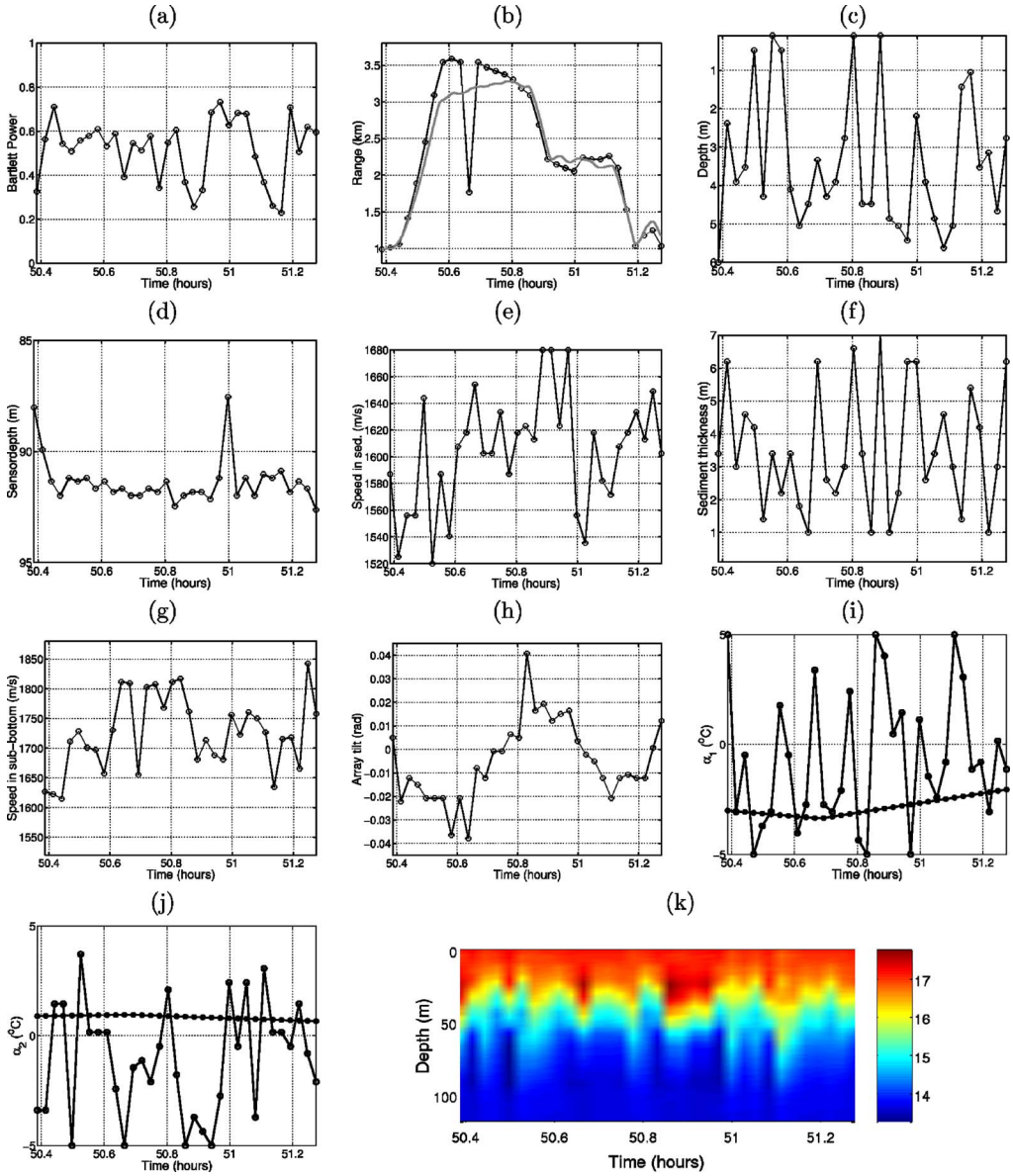


FIG. 12. (Color online) Focalization results for event 6: Bartlett power (a), source range (b) [the continuous line is the GPS measured source–receiver range], source depth (c), receiver depth (d), sediment compressional speed (e), sediment thickness (f), sub-bottom compressional speed (g), VLA tilt (h), EOF coefficient α_1 (i), EOF coefficient α_2 (j) [filled dotted lines are the XBT measured data projected onto the respective EOFs], and reconstructed temperature (k).

out the run at 359 and 719 Hz, where one might be a harmonic of the other, as well as other high coherence bins with some persistence at 490, 498, and 542 Hz. Actually, instead of setting γ_0, and in order to limit the computational complexity of the problem, a fixed number of frequency bins were selected at each time slot according to the maximum values of (6). The inversion results are shown in Fig. 12: from (a) to (j) are individual parameter estimates, while plot (k) shows the water column temperature reconstruction based on the EOF linear combination of parameter estimates (i) and (j). Plot (b) shows the estimated source range together with the GPS measured source range (continuous line). At first glance from the model fit indicator (Bartlett power) the result is poor, since it is always below 0.8; however, source range, which is one of the leading parameters, shows values in coincidence with the GPS measured source range while source depth is highly incoherent within the 0–6-m depth interval; and finally the reconstructed temperature [plot (k)] appears to be too variable for such a small time interval (slightly over 45 min). Looking more in detail, and comparing plot (a) of Fig. 5 with plots (a)–(c) of Fig. 12, the following conclusions can be drawn: (i) for $50.42 \leqslant \text{time} \leqslant 50.57$, ship speed increases steeply to 9 kn, while heading off from the VLA. Range variation is about 4.6 m/s, which may cause a violation of the stationary assumption during the averaging time. Source range estimation error progressively increases as the ship reaches the longest range point and then remains high during part of the first loop trajectory at an approximate constant range of 3.2 km; this erroneous source range estimate is almost certainly due to the environmental water depth mismatch in this portion of the track both along the NE track (event 5) and part of the trajectory to the NW track (see Fig.2).

After the end of the first loop the estimated source range perfectly matches the GPS curve. As mentioned above, source depth [plot (c)] is highly variable in the interval 0–6-m depth, which is understood to be due to the nature of the emitted signal (ship noise) radiated from a structure extending below but also on or above the sea surface. Receiving array estimates shown in plots (d) and (h) for array depth and tilt, respectively, are in agreement with the expected values. Notice the interesting behavior of array tilt that varies from −0.03 to +0.03 almost linearly during the side-looking view 45 deg change when running from the NE to the NW leg and then back to the NE leg at the end of the run. Concerning the seafloor properties, one can say that their estimates are confined to relatively short time intervals at periods when the vessel is steaming at her maximum speed, which somehow justifies their high variability as seen on plots (e) to (g). EOF coefficients α_1 and α_2 are shown in plots (i) and (j), respectively. The estimated values are highly variable within the search interval, which is believed to be due to the highly variable and extremely low number of frequencies available in the ship-radiated noise spectrum, associated with ship's acceleration and deceleration during maneuvering. As a final comment on Fig. 12, the reconstructed water temperature—plot (k)—suffers both from poor estimation and ship variability. A different way of looking at the results is to plot histograms of the estimates as shown in Fig. 13, where it can be seen that α_1's most frequent estimate is −3 and that of α_2 is 0, which are values compatible with those measured with the XBT during that period of time [filled dots curve on plots (i) and (j)].

VI. DISCUSSION AND FINAL CONCLUSIONS

OAT is an appealing technique for remote monitoring of the ocean volume. One of the basic principles of OAT is that both source(s) and receiver(s) are under control of the experimentalist, that is, the emitted source signal and the source–receiver geometry is known (with some degree of precision) at all times during the observation window. In passive tomography the control of the source is relaxed, in order to be able to take advantage of possible sources of opportunity passing within acoustic range from the receiver(s). Although passive tomography is very appealing for the ease of application, its practical implementation is extremely challenging and its feasibility remains to be proved.

This paper shows the tomographic inversion results obtained on a dataset obtained during the INTIFANTE'00 sea trial aiming at proving the feasibility of passive tomography. The results are presented in increasing order of difficulty for applying BOAT with a deterministic moving source in a range-independent environment, a pseudorandom source in a range-dependent environment, and finally using the noise radiated by a fast moving ship in a partially range-dependent environment. The challenge is represented by the fact that during the various phases of the processing the *a priori* knowledge about the source is progressively relaxed, leading to a situation close to that encountered in plain passive tomography.

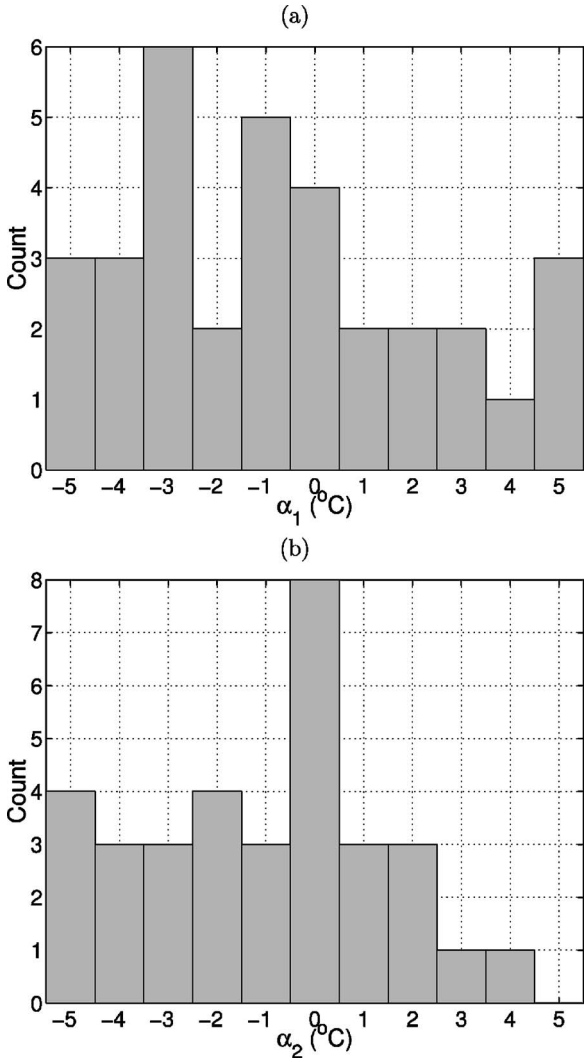


FIG. 13. INTIFANTE'00 sea trial, event 6: histograms of the EOF coefficients estimates for α_1 (a) and α_2 (b).

In a first data set it is proved that a moving source at an unknown location emitting a deterministic unknown signal over a range-independent environment can be used for ocean tomography adapting to the environment in time. Estimates of the various environmental and geometric parameters are consistent with the expected values. Focalization was demonstrated to represent the tool of choice, accounting for the unknown geometric and environmental parameters. In a second data set a moving source at an unknown location was emitting a PRN sequence used for determining the sound-speed structure as well as other geometrical and geoacoustic parameters over a range-dependent environment. The results were also consistent with expectations and demonstrated that the employed methods can operate, although with increasing difficulty, in range-dependent environments with both deterministic and pseudorandom noise source signals. Most importantly, it was shown that high-ranking parameters such as source range and depth together with the Bartlett fit could be used as indicators of the estimates' validity throughout the run. Finally, a third data set, where the controlled sound source was replaced by the ship itself moving at high speed in a series of concentric loops around the receiving array,

showed that during some "clear" periods credible environmental estimates were obtained with the source ship correctly positioned and presumed correct environmental parameters.

During the analysis of this data set severe concerns were raised regarding the proper account for the vertical line array (VLA) geometry into the processor. Based on real-time measurements of the array tilt and sensor depth, the VLA was suspected to be responsible for several data mismatches during the inversion process. Those mismatches were seen to happen in conjunction with tidal movements at the array location. Another concern relates to the useful bandwidth of the radiated ship noise for environmental inversion. Although that concern was partially mitigated by the frequency selection based on the signal's short time coherence, a doubt remains whether real cargo ships at cruising speeds do radiate enough bandwidth sufficiently loud for the purpose of BOAT. The authors believe that these results demonstrate that blind ocean acoustic tomography is feasible and worth further development.

ACKNOWLEDGMENTS

The authors would like to thank the Instituto Hidrográfico, Lisbon (Portugal) for the organization of the INTIFANTE'00 sea trial and E. Muzi from NURC for its participation in the sea trial. This work was partially financed by FCT, Portugal, under Contract ATOMS, PD-CTM/P/MAR/15296/1999 and Project TOMPACO from CNR (Italy). The comments and suggestions of the two anonymous reviewers are greatly appreciated for their contribution to the clarity of the paper.

[1]M. J. Buckingham and S. A. S. Jones, "A new shallow-ocean technique for determining the critical angle of the seabed from the vertical directionality of the ambient noise in the water column," J. Acoust. Soc. Am. **81**(4), 938–946 (1987).

[2]M. J. Buckingham and J. R. Potter, "Acoustic daylight imaging: vision in the ocean," GSA Today **4**, 97–102 (1994).

[3]N. M. Carbone, G. B. Deane, and M. J. Buckingham, "Estimating the compressional and shear wave speeds of a shallow water seabed from the vertical coherence of ambient noise in the water column," J. Acoust. Soc. Am. **103**(2), 801–813 (1998).

[4]C. H. Harrison and A. Baldacci, "Bottom reflection properties by inversion of ambient noise," in *Proc. Sixth of European Conf. on Underwater Acoust.*, ECUA'02, pages 471–476, Gdansk, Poland, June 2002.

[5]C. H. Harrison, "The influence of noise and coherence fluctuations on a new geo-acoustic inversion technique," in *Impact of Littoral Environment Variability on Acoustic Predictions and Sonar Performance*, edited by N. G. Pace and F. B. Jensen (Kluwer, Dordrecht, 2002), pp. 139–146.

[6]C. H. Harrison and D. G. Simons, "Geoacoustic inversion of ambient noise: A simple method," J. Acoust. Soc. Am. **4**(112), 1377–1389 (2002).

[7]M. J. Buckingham, E. M. Giddens, J. B. Pompa, F. Simonet, and T. R. Hahn, "A light aircraft as a source of sound for performing geo-acoustic inversion of the sea bed," in *Proc. of Sixth European Conf. on Underwater Acoust.*, ECUA'02, pages 465–470, Gdansk, Poland, June 2002.

[8]J. A. Cadzow, "Blind deconvolution via cumulant extrema," IEEE Signal Process. Mag. **13**, 24–42 (1996).

[9]W. A. Gardner, "A new method of channel identification," IEEE Trans. Commun. **39**, 813–817 (1991).

[10]L. Tong, G. Xu, and T. Kailath, "A new approach to blind identification and equalization of multipath channels," in *Proc. of the 25th Asilomar Conf. on Signals, Systems and Computers*, 1991.

[11]M. K. Broadhead, "Broadband source signature extraction from underwater acoustics data with sparse environmental information," J. Acoust. Soc. Am. **97**, 1322–1325 (1995).

[12]M. K. Broadhead and L. A. Pflug, "Performance of some sparseness criterion blind deconvolution methods in the presence of noise," J. Acoust. Soc. Am. **107**, 885–893 (2000).

[13]P. Gerstoft, "Inversion of seismoacoustic data using genetic algorithms and *a posteriori* probability distributions," J. Acoust. Soc. Am. **95**(2), 770–782 (1994).

[14]J. -P. Hermand and P. Gerstoft, "Inversion of broad-band multitone acoustic data from the yellow shark summer experiments," IEEE J. Ocean. Eng. **21**(4), 324–364 (1996).

[15]J. -P. Hermand, "Broad-band geoacoustic inversion in shallow water from waveguide impulse response measurements on a single hydrophone: Theory and experimental results," IEEE J. Ocean. Eng. **24**(1), 41–66 (1999).

[16]M. Siderius, P. L. Nielsen, J. Sellschopp, M. Snellen, and D. Simons, "Experimental study of geo-acoustic inversion uncertainty due to ocean sound-speed fluctuations," J. Acoust. Soc. Am. **110**(2), 769–781 (2001).

[17]C. -S. Chiu, J. H. Miller, W. Denner, and J. F. Lynch, "Forward modeling of the barents sea tomography vertical line array data and inversion highlights," in *Full Field Inversion Methods in Ocean and Seismo-Acoustics*, edited by O. Diachok *et al.* (Kluwer Academic, Netherlands, 1995), pp. 237–242.

[18]C. Soares, A. Waldhorst, and S. M. Jesus, "Matched field processing: Environmental focusing and source tracking with application to the north elba data set," in *Proc. of the Oceans'99 MTS/IEEE Conference*, pages 1598–1602, Seattle, Washington, 13–16 September 1999.

[19]P. Gerstoft and D. Gingras, "Parameter estimation using multi-frequency range-dependent acoustic data in shallow water," J. Acoust. Soc. Am. **99**(5), 2839–2850 (1996).

[20]C. Soares, M. Siderius, and S. M. Jesus, "Source localization in a time-varying ocean waveguide," J. Acoust. Soc. Am. **112**(5), 1879–1889 (2002).

[21]M. D. Collins and W. A. Kuperman, "Focalization: Environmental focusing and source localization," J. Acoust. Soc. Am. **90**(3), 1410–1422 (1991).

[22]C. Soares and S. M. Jesus, "Broadband matched-field processing: Coherent vs incoherent approaches," J. Acoust. Soc. Am. **113**(5), 2587–2598 (2003).

[23]S. M. Jesus and C. Soares, "Blind ocean acoustic tomography with source spectrum estimation," in *Proc. Int. Conf. on Theoretical and Computational Acoustic*, Honolulu, Hawaii, August 2003.

[24]S. M. Jesus, E. Coelho, J. Onofre, P. Picco, C. Soares, and C. Lopes, "The Intifante'00 sea trial: Preliminary source localization and ocean tomography data analysis," in *Proc. of the MTS/IEEE Oceans 2001*, Honolulu, Hawaii, 5–8 November 2001.

[25]O. C. Rodríguez, S. Jesus, Y. Stephan, X. Démoulin, M. Porter, and E. Coelho, "Internal tide acoustic tomography: Reliability of the normal modes expansion as a possible basis for solving the inverse problem," in *Proc. of the 4th. European Conference on Underwater Acoustics*, pages 587–592, Rome, Italy, 21–25, September 1998.

[26]C. Soares, "Matched-field processing: Acoustic focalization with data taken in a shallow water area of the Strait of Sicily," MSC Report, SiPLAB-FCT, University of Algarve, Faro, Portugal, March 2001.

[27]C. Soares, M. Siderius, and S. M. Jesus, "High frequency source localization in the Strait of Sicily," in *Proc. of the MTS/IEEE Oceans 2001*, Honolulu, Hawaii, November 2001.

[28]T. Fassbender, 1995, "Erweiterte genetische algorithmen zur globalen optimierung multi-modaler funktionen, Diplomarbeit, Ruhr-Universität Bochum 1995.

[29]C. M. Ferla, M. B. Porter, and F. B. Jensen, *C-SNAP: Coupled SACLANTCEN normal mode propagation loss model*. La Spezia, Italy.

[30]S. M. Jesus and C. Soares, "Tomografia passiva costiera, inversion results with active data - phase 2," Internal Report Rep. 06/01, SiPLAB/CINTAL, Universidade do Algarve, Faro, Portugal, December 2001.

[31]S. M. Jesus, C. Soares, J. Onofre, E. Coelho, and P. Picco, "Experimental testing of the blind ocean acoustic tomography concept," in Ref. 5, pp. 433–440.

Cost-effective assembly of a basic fiber-optic hydrophone for measurement of high-amplitude therapeutic ultrasound fields

Jessica E. Parsons[a)]
Department of Biomedical Engineering, University of Michigan, Ann Arbor, Michigan 48109

Charles A. Cain
Departments of Biomedical Engineering and Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109

J. Brian Fowlkes
Departments of Biomedical Engineering and Radiology, University of Michigan, Ann Arbor, Michigan 48109

(Received 20 August 2005; revised 16 December 2005; accepted 16 December 2005)

Design considerations, assembly details, and operating procedures of one version of a cost-effective basic fiber-optic probe hydrophone (FOPH) are described in order to convey practical information to groups interested in constructing a similar device. The use of fiber optic hydrophones can overcome some of the limitations associated with traditional polyvinylidene difluoride (PVDF) hydrophones for calibration of acoustic fields. Compared to standard PVDF hydrophones, FOPH systems generally have larger bandwidths, enhanced spatial resolution, reduced directionality, and greater immunity to electromagnetic interference, though they can be limited by significantly lower sensitivities. The FOPH system presently described employs a 100-μm multimode optical fiber as the sensing element and incorporates a 1-W laser diode module, 2×2 optical coupler, and general-purpose 50-MHz silicon *p-i-n* photodetector. Wave forms generated using the FOPH system and a reference PVDF hydrophone are compared, and intrinsic and substitution methods for calibrating the FOPH system are discussed. The voltage-to-pressure transfer factor is approximately 0.8 mV/MPa (-302 dB re 1 V/μPa), though straightforward modifications to the optical components in the FOPH system are discussed that can significantly increase this value. Recommendations are presented to guide the choice of optical components and to provide practical insight into the routine usage of the FOPH device.
[DOI: 10.1121/1.2166708]

PACS number(s): 43.35.Yb, 43.38.Zp, 43.58.Vb [TDM] Pages: 1432–1440

I. INTRODUCTION

The use of fiber optic probe hydrophones (FOPH) for acoustic field calibration was first proposed by Phillips[1] and Staudenraus and Eisenmenger[2] and has since been developed and characterized extensively. Herein, the present authors discuss their practical experiences using one method to assemble a cost-effective functional FOPH system for the calibration of high-amplitude therapeutic ultrasound fields using widely available optical components from the telecommunications industry (approximate cost: $5500, not including a precision fiber cleaving device). The present authors' purpose is to communicate practical information gathered from trial-and-error experiences and to suggest design improvements that may assist other groups in the construction and routine usage of a FOPH device.

In brief, the FOPH system measures the pressure variations produced by an acoustic source by launching light from a laser module into a single-mode or multimode optical fiber positioned in the acoustic focus and monitoring the temporal fluctuations in the reflected light level. The proportion of incident light that is reflected at the fiber endface is modulated by the pressure field's dynamic effect on the refractive index of the host fluid. By monitoring the modulated light intensity, a temporally varying voltage signal is obtained that is directly proportional to the acoustic pressure. This type of FOPH device offers several widely-recognized advantages over standard polyvinylidene difluoride (PVDF) membrane or needle hydrophones.[2–5] These features include the FOPH system's enhanced spatial resolution, broader bandwidth, greater immunity to electromagnetic interference, and ability to register higher pressure amplitudes compared to PVDF hydrophones. Furthermore, the FOPH system can be intrinsically calibrated without reliance on a reference standard, and the sensor tip can be quickly repaired after exposure to high-amplitude fields and associated cavitation activity that would irreparably damage PVDF hydrophones. These advantages are tempered, however, by a significant reduction in sensitivity with respect to PVDF hydrophones due to an inherently low signal-to-noise ratio. Without additional enhancements to the system, this fundamental sensitivity limitation restricts the practicality of the FOPH device discussed herein to use with relatively high-amplitude acoustic fields ($\gtrsim$0.9 MPa). Therefore, the FOPH device described by the present authors is of primary interest for calibration of trans-

[a)]Electronic mail: jeparson@engin.umich.edu

ducers operating at therapeutic pressure levels, though suggested improvements and references to other configurations are discussed that may increase both its sensitivity and bandwidth to enable its use at the amplitudes and frequencies typically associated with imaging transducers.

II. PRINCIPLE OF OPERATION AND INTRINSIC CALIBRATION

The operating principles and theoretical treatments of the FOPH system have been described previously.[2–4,6] In summary, the FOPH device produces measurements of pressure fields by exploiting the piezo-optic effect, in which the propagating pressure wave fronts temporally affect the densities and, hence, the indices of refraction of both the fluid and the fiber endface inserted into the fluid. However, due to the low compressibility of the fiber optic material with respect to that of the host fluid (e.g., water), the change in refractive index of the fiber is typically neglected in FOPH systems. Light emitted by a laser diode module propagates to the fiber endface via an optical coupler (either 2×2 or 1×2 configuration) and is then partially reflected back into the fiber at the fiber-water interface, as governed by the presiding refractive index of the water. The refractive index of the host water in proximity to the fiber endface fluctuates due to the piezo-optic effect and therefore modulates the light intensity that is reflected back into the fiber and subsequently directed to a photodetector via the optical coupler. The voltage output of the photodetector is converted into the equivalent acoustic pressure wave form via substitution methods using reference standards such as another calibrated hydrophone or by intrinsic parameter-based calibration of the FOPH system (described subsequently).

The piezo-optic relationship between the acoustic pressure variation and the induced change in refractive index that governs the FOPH system can be determined via application of a modified isentropic version of the Tait equation,[2,3,7] which describes the pressure-density dependence, in conjunction with the Gladstone–Dale relation,[2,3,8–10] which describes the density-refractive index dependence. The pertinent version of the Tait equation, valid in water up to 1 GPa, can be represented by

$$\frac{P_0 + p(t) + Q}{\rho(t)^{\gamma}} = K_1, \tag{1}$$

where K_1 is a constant, P_0 is the ambient pressure, $p(t)$ is the acoustic pressure, $\rho(t)$ is the density of water, and Q and γ are fit parameters determined to be 295.5 MPa and 7.44, respectively, under standard conditions of $T=20\ ^{\circ}\mathrm{C}$, $P_0=100$ kPa, and $\rho_0=1000\ \mathrm{kg/m^3}$.[2,7] Additionally, the Gladstone–Dale model, valid for compressional pressures up to approximately 500 MPa, relates the density and refractive index of water by

$$\frac{n_{\mathrm{water}}(t) - 1}{\rho(t)} = K_2, \tag{2}$$

where K_2 is again a constant.[2,8–10] Simultaneous solution of Eqs. (1) and (2) leads to determination the pressure-dependent change in the refractive index of water that arises in response to the acoustic field. The change in refractive index of water varies nearly linearly with the applied pressure in the pressure range of interest (approximately 10^{-2} MPa to 10^{2} MPa),[4] which allows for the definition of the piezo-optic constant, $\partial n/\partial p$. The empirically and theoretically determined values of $\partial n/\partial p$ for water range from $1.32\times10^{-4}\ \mathrm{MPa^{-1}}$ to $1.66\times10^{-4}\ \mathrm{MPa^{-1}}$.[1,2,4,10,11] In comparison, the measured piezo-optic constant for the optical fiber material is typically two orders of magnitude smaller than that of water (e.g., $\partial n_{\mathrm{silica}}/\partial p=5\times10^{-6}\ \mathrm{MPa^{-1}}$),[4,12] justifying the common practice of neglecting the change of refractive index of the fiber during ultrasonic field measurements using the FOPH system.

A straightforward technique has been described by Huber *et al.*[6] to intrinsically determine the FOPH voltage-to-pressure transfer factor by incorporating the refractive index of the optical fiber and the piezo-optic constant for water reported in the literature. The transfer factor S as a ratio of voltage change (ΔV) to pressure change (Δp) can be defined as:

$$S = \frac{\Delta V}{\Delta p} = \frac{\Delta V}{\Delta n}\cdot\frac{\Delta n}{\Delta p}, \tag{3}$$

where $\Delta n/\Delta p \approx \partial n/\partial p$ under linear conditions. Therefore, simple measurement of the dc voltage change resulting from the static reflection that occurs when the cleaved fiber endface is immersed into the water tank constitutes a direct means by which to determine the FOPH transfer factor. This transfer factor is then used to obtain corresponding pressure amplitudes from the voltage wave form produced by the photodetector. It should be noted that in addition to calibration methods relying on the use of the piezo-optic constant of water, alternative techniques for parameter-based intrinsic calibration of the FOPH system are discussed by Staudenraus and Eisenmenger[2] and Wurster *et al.*[3]

III. MATERIALS AND METHODS

The FOPH system described herein is constructed using commercially available standard optical components including a laser diode module with 1 W (nominal) optical power, a 2×2 (four-arm) fiber coupler, a heatsink that serves to stabilize the diode temperature as well as to provide electrical connection to the diode driver, and a simple silicon *p-i-n* photodiode to monitor light fluctuation. Figure 1 summarizes the components and their layout in the FOPH system.

A. Fiber-coupled ("pigtailed") laser diode

In order to maximize the ease of construction, errors associated with collimation procedures to couple light into the optical fiber were avoided through the use of a "pigtailed" laser diode. These diodes offer the distinct advantage of having the optical fiber coupled directly to the laser diode in a hermetic package during manufacturing, thereby eliminating the need for collimation lenses or other optical instrumentation. There is an associated coupling loss when using such a pigtailed diode (approximately 30%), but the use of a 1.5 W diode chip assures that adequate optical power is

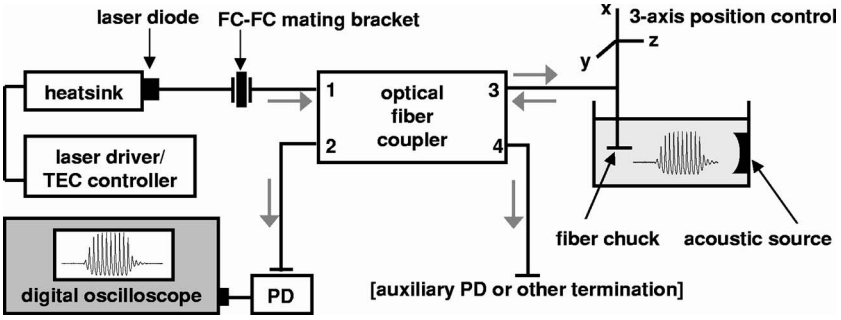


FIG. 1. Components comprising the FOPH system. A hermetically sealed, fiber-coupled laser diode with FC termination is mounted directly to a heatsink device, which provides electrical connection of the laser driver/temperature controller console to both the integrated thermoelectric cooler and the laser diode. The laser diode pigtail is coupled to Port 1 of the optical coupler via a FC-FC mating bracket. The fiber attached to Port 3 of the coupler is left bare with no connector so that the endface can be cleaved and positioned in the calibration tank using a fiber chuck that is mounted to a three-axis motorized positioning system. The signal reflected at the fiber endface is directed via the coupler to a photodetector connected to Port 2, and the resultant voltage is measured on an oscilloscope. The coupler's fourth port can be attached to another photodetector for noise suppression purposes or can be otherwise terminated to minimize reflections. The grey arrows indicate the direction of travel of the optical signal during operation.

launched into the coupler from the pigtailed diode despite this coupling loss. The laser diode employed (QPhotonics, Chesapeake, VA, model QLD-850-1000M) has a wavelength of 850±5 nm, a threshold current of 300 mA, and is typically operated during acoustic measurements using a forward current of 1350 mA. It is housed in an eight-pin TO3-type package with an embedded thermoelectric microcooler and mounts directly to a heatsink device discussed subsequently. The fiber pigtail is ~0.5 m in length, and is composed of graded-index multimode glass fiber with core/cladding diameters of 100/125 μm, respectively. The fiber pigtail is terminated with a standard fiber connector (FC)-type optical connector.

B. Heatsink and laser driver

The laser diode unit is mounted on an air-cooled heatsink (Laser Drive, Gibsonia, PA, model LDI-800-H) that serves to link the diode and thermoelectric cooler to the diode driver (Laser Drive, Gibsonia, PA, model LDI-805). The heatsink is connected via manufacturer-supplied cabling to the laser diode driver, which consists of a microprocessor-controlled low-noise current source and precision temperature controller with IEEE 488.2 communications capability.

C. Fiber optic coupler and photodetector

The fiber optic coupler is a 2×2 (four arm), 3-dB (50:50 coupling ratio), bidirectional fused device consisting of two input leads and two output leads composed of 100/140 μm SpecTran graded-index multimode glass fiber with 900 μm Hytrel jacketing (Gould Fiber Optics, Millersville, MD, model 15-22500-50-66640, package type 42). The measured insertion losses range from 3.27–3.55 dB (including connector losses), while the directivity (near-end crosstalk) is specified to be $\geqslant$40 dB across the input leads. Each of the input leads and one output lead are terminated with FC connectors, while the remaining output lead is the test lead submerged during measurements and therefore has no connector. Each of the leads with connectors has a standard length (21 cm), while the test lead is 300 cm (maximum length available from the manufacturer). When the FOPH system is not in use, the test lead is wound around a protective fiber storage spool (Thorlabs, Newton, NJ, model PCS1-X) that is mounted to polycarbonate housing constructed in house. The laser diode pigtail is connected to one of the input leads on the fiber optic coupler via a FC-FC mating bracket (Thorlabs, Newton, NJ, model FCB1 or ADAFC1). The loss associated with the FC-FC mating bracket is specified to be 0.1–0.2 dB. The second input lead is connected to a standard fixed-gain silicon *p-i-n* photodetector (Thorlabs, Newton, NJ, model PDA-155) via a simple mount and adapter system that allows the FC connector on the second input lead to screw into the photodetector housing (Thorlabs, Newton, NJ, model SM1FC, SM1T2, PH2-ST, BA2, and TR2, used in conjunction). The photodetector has a dc-50 MHz bandwidth, a fixed 5×10^3 V/A transimpedance gain, 0.325 A/W responsivity at 850 nm, and a noise-equivalent power of 4×10^{-11} W/$\sqrt{}$Hz (max). No external preamplifier is used beyond the transimpedance gain provided by the photodetector.

Light launched into the first input lead from the laser diode module is attenuated by ~3 dB as it is split via the optical coupler and directed to both the test lead and the second output lead. When the test lead is placed in an acoustic field, the resultant temporally varying reflections generated at the fiber endface are directed back through the coupler, again being attenuated by ~3 dB, to each of the two input leads. In this manner, the photodetector voltage monitored on an oscilloscope (LeCroy, Chestnut Ridge, NY, model 9384) produces a representation of the pressure wave form. It should be noted that in any FOPH device using similar optical couplers, light propagating back through the coupler due to reflection at the test lead/water interface will not only be directed to the photodetector but also into the input lead to which the laser diode is connected. This does not pose a significant risk to the integrity of the laser, however, because the optical power reflected back into the coupler at the fiber/water interface is on the order of 0.3% of that incident on the endface and is further attenuated by an additional ~3 dB as it passes back through the coupler. The optical power reflected into the coupler lead connected to the laser diode is therefore miniscule, typically less than 0.8 mW. It is also important to note that in the present sys-

tem a conscious choice was made to avoid additional losses that would be associated with a removable connection between the coupler housing and the test lead by using a coupler that is permanently integrated with the test lead. The disadvantage of this decision is that when the test lead length has been depleted due to repeated cleaving of the fiber, an entire new coupler must be supplied. However, the cost of the coupler device is modest and therefore its periodic replacement does not pose a significant burden to the user.

D. Usage options for second coupler output lead

The incorporation of a 2×2 optical coupler affords several possibilities for the use of the second output lead, though this lead is not directly required for operation of the FOPH system. The simplest of these options is to not employ the second output lead and instead to terminate it either by submerging the lead in a small water bath or by using a small amount of index-matching gel. In the present authors' experience, terminating the second output lead in a vial containing distilled water has provided acceptable results. In these cases, the FC connector on the second output lead is not necessary and the user may opt to remove it. Both of these termination methods provide conditions under which the reflection at the unused output lead endface is miniscule. More sophisticated uses of the second output lead certainly exist, however, including the use of a second photodetector attached to the unused output lead to monitor the optical power throughout a calibration procedure or to measure the noise level arising from the optical source. This noise signal can be scaled and subtracted from the pressure signal as a simple noise suppression technique.[2,3] Note that the placement of a photodetector on the second coupler output lead may require the use of a neutral-density optical filter between the lead and the photodetector to prevent damage to the photodiode (e.g., if a duplicate photodetector to the one specified herein is placed on the second output lead, the photodiode damage threshold of 100 mW will be exceeded under normal use).

E. Methods of calibration

Calibration of the FOPH system can be generally accomplished using two methods. Substitution calibration involves comparing wave forms generated by an acoustic source using identical input conditions recorded using both a calibrated reference hydrophone and the FOPH system. This substitution method allows the user to determine the appropriate transfer factor for the FOPH without any detailed knowledge of the optical parameters of the FOPH device but does require reliance on a reference standard and can be time consuming. Note that this method of substitution calibration should be ideally performed using low-amplitude linear wave forms where the compressional and rarefactional pressures are similar. In the event that wave forms generated for comparison are not symmetric, it is more appropriate to compare the voltages corresponding to the rarefactional pressures only, as discrepancies in the voltage amplitudes corresponding to compressional pressures readily arise due to the greater bandwidth and smaller active aperture of the FOPH system compared to common PVDF membrane or needle hydrophones. However, while it is less problematic to compare voltages corresponding to peak rarefactional pressures, it should be noted that this approach is itself only an approximate method for calibrating the FOPH system using substitution techniques. This is due in part to the fact that rarefactional pressures recorded using the reference PVDF membrane hydrophone can be distorted due to frequency-dependent variations in its transfer function. Furthermore, the discrepancy in active aperture dimensions between the reference PVDF hydrophone and the FOPH system becomes increasingly influential as the ultrasonic wavelength is increased and may lead to complications in comparing the rarefactional pressures registered. A comprehensive discussion of these issues has been provided by Wilkens and Koch.[13] Therefore, proper calibration of the FOPH system using substitution techniques may require spectral evaluation and/or the application of appropriate spatial averaging corrections.

As an alternative to substitution calibration that does not require the use of reference standards, the FOPH system can be intrinsically calibrated by measuring the static dc reflectance in water and applying Eq. (3) above. Routine measurement of the static dc reflectance should be performed each time the FOPH device is used, and following any recleaving of the fiber tip. This simple measurement allows for minor adjustments in the transfer factor to be made to account for any instability in the optical power output of the laser diode and/or variation in fiber cleaving/stripping techniques that may arise (cf. Sec. V A).

F. Acoustic measurement procedure

In order to prepare the fiber for use during acoustic calibration procedures, the fiber is first stripped of its outer jacket and cladding using a stripping device (Thorlabs, Newton, NJ, model T08SG40). The fiber is then inserted through the barrel of a fiber optic chuck (Newport, Irvine, CA, model FPH-DJ) and subsequently cleaved to produce a flat endface normal to the fiber axis using a high-precision cleaving tool (Thorlabs, Newton, NJ, model XL410). Once the fiber has been prepared, it is situated so that approximately 1 cm of the bare glass protrudes beyond the chuck. The chuck is then mounted in the calibration tank and the FOPH system is activated. The laser diode is operated under continuous-wave conditions throughout the calibration procedure, typically using a forward current of $\sim$1350 mA. The thermoelectric cooler and heatsink maintain the diode package temperature at 25 °C throughout the procedure using up to 3.5 A for temperature stabilization.

Due to the relatively low FOPH sensitivity, wave form averaging is usually necessary to observe the time-varying voltage signal adequately during measurements of lower-pressure amplitudes (e.g., 64–256 point averaging for $P_r \leq 5$ MPa). The amount of averaging required decreases as the pressure amplitude increases and, generally, minimal averaging (e.g., 16 point) is adequate to obtain acceptable signals above rarefactional pressures of $\sim$10 MPa. Note that the voltage wave forms generated by the FOPH device must be inverted prior to converting to pressures, due to the fact that

positive (compressional) pressures temporarily increase the fluid density at the fiber tip, thereby reducing the refractive index mismatch at the fiber-fluid interface and causing a *negative* change in the reflected light level.

In the event of cavitation inception at the fiber endface due to the application of large-amplitude acoustic fields, a rapid rise in reflectance is observed at the photodetector with a concomitant loss of wave form integrity. It is usually obvious to the operator when cavitation at the fiber tip has occurred, and this event can be readily verified by remeasuring the dc reflectance to determine whether it has risen substantially above its initial value. If the fiber tip has indeed been corrupted, it can then be removed from the measurement system, restripped and recleaved to produce a new endface, and repositioned at the same location in the water tank to continue measurements.

G. Comments on the use of 1×2 fiber optic couplers

The preceding discussion indicates that the use of the second output lead on the fiber optic coupler is optional and is not necessary for acceptable FOPH performance. Therefore, the use of 1×2 fiber optic couplers is also suitable, as other groups have demonstrated.[6] However, in the present authors' experience, technical aspects associated with the manufacturing process of 1×2 couplers degraded their long-term performance in the current FOPH system. These technical complications arise from the fact that for any FOPH system, the fiber optic coupler must route light launched from the laser source to the test lead and then redirect the reflected light through the coupler again to the lead to which the photodetector is connected. Therefore, the laser diode lead and the photodetector lead must be on opposite sides of the coupler from the test lead. In 1×2 fiber optic couplers, which have one input lead that bifurcates into two output leads, this requires that the device be used in reverse of its typical orientation, i.e., that the light must be launched into one of the *output* leads to travel to the test lead on the *input* terminal of the device. The supplier of the coupler used in the current FOPH system (Gould Fiber Optics, Millersville, MD) manufactures its 1×2 couplers from 2×2 couplers by truncating the fourth lead internally within the device and terminating the remnant of the fourth lead with an epoxy bead. Using the device in the reverse orientation allows high-amplitude light to be incident on the epoxy bead, which was observed to char over time as a result. This effect led to an overall system noise increase and loss of hydrophone sensitivity. The present authors are unaware whether this is a manufacturer-specific issue and therefore recommend the general use of 2×2 optical couplers to eliminate this complication and to accommodate the additional refinements discussed above that are available using the second coupler output lead. It should be noted that an ideal alternative to the use of 1×2 or 2×2 optical couplers for light routing requirements would be provided by a three-port optical circulator; however, these devices are less readily available than optical couplers for multimode fiber or near-IR applications and are significantly more expensive.

IV. RESULTS

A. Determination of FOPH transfer factors

The appropriate transfer factor was obtained first via substitution calibration by comparing peak rarefactional pressures of linear wave forms generated by a 788-kHz spherical shell transducer (Piezo Technologies, Etalon Division, Lebanon, IN) driven using identical excitation conditions. These wave forms were recorded using a reference PVDF bilaminar shielded membrane hydrophone with an active element diameter of 1 mm (GEC Marconi Research Center, Essex, UK; calibrated by Sonic Consulting, Wyndmoor, PA) and, subsequently, the FOPH system. Wave forms were sampled at $f_s=500$ MHz using a digital storage oscilloscope. The calibration performed using this substitution technique yielded an average FOPH transfer factor of 0.78 ± 0.04 mV/MPa [−302.1 dB re 1 V/μPa; mean ± standard deviation (SD), $N=6$] without the use of external preamplifiers or noise suppression/filtering. Second, the transfer factor was computed intrinsically using the group refractive index of the optical fiber specified at $n_f=1.497$, the refractive index of water at 850 nm($n_w=1.329$),[14] the static dc reflectances measured at the laser diode operating current used during wave form collection (785 ± 51 mV at 1350 mA, mean ± SD, $N=6$), and the piezo-optic constant for water determined according to the method used by Phillips[1] (1.66×10^{-4} MPa^{-1} for $\lambda=850$ nm). Using these parameters, the average intrinsic transfer factor was determined to be 0.775 ± 0.05 mV/MPa (−302.2 dB re 1 V/μPa; mean ± SD, $N=6$), which is in excellent agreement with the average value determined by substitution calibration. Factors affecting the precision of these transfer factor measurements and the accuracy of the resultant pressure amplitudes obtained using them are discussed subsequently (cf. Sec. V A).

B. Wave form examples

Figure 2 provides a comparison of wave forms measured by the FOPH system (*right panels*) and the PVDF membrane hydrophone (*left panels*). In all panels, the acoustic source was the 788-kHz spherical shell transducer described above emitting a ten-cycle pulse. The transducer was driven using identical excitation conditions in Figs. 2(a) and 2(b), and was then driven at identical higher-amplitude conditions in Figs. 2(c) and 2(d) in order to elicit shocking. The FOPH wave forms were averaged prior to recording to improve the signal-to-noise ratio, but were not filtered or externally amplified beyond the conversion gain provided by the photodetector. Though the FOPH wave forms are generally noisier than the wave forms recorded with the PVDF membrane hydrophone, comparison of Figs. 2(c) and 2(d) indicates that the enhanced bandwidth and smaller active aperture of the FOPH system allows it to more accurately measure the peak compressional pressures present when shocked wave forms develop. This feature may be of critical importance during measurement of highly nonlinear therapeutic ultrasound fields.

Figure 3 shows two wave forms of limited nonlinearity generated under identical excitation conditions using a 750-kHz annular array (Imasonic, S.A., Besançon, France)

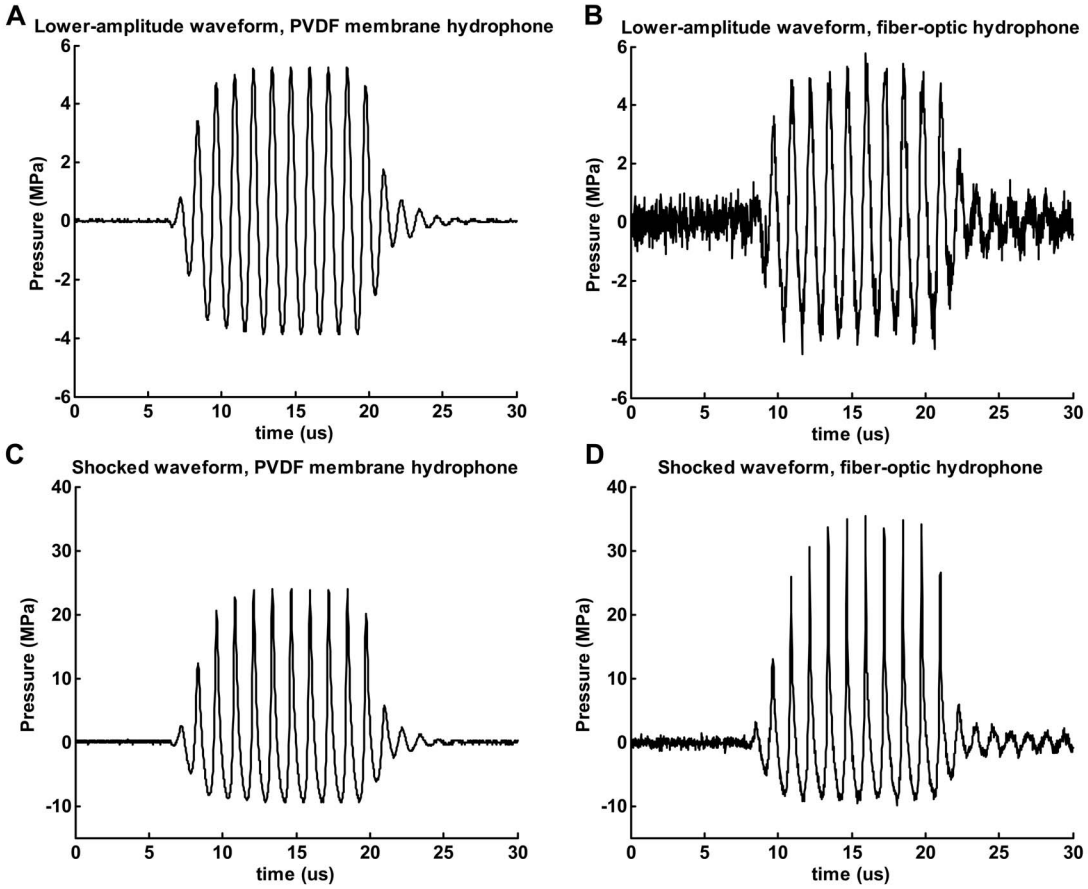


FIG. 2. Pressure wave forms collected using a PVDF membrane hydrophone with 1-mm active diameter (*left panels*), compared to those collected using the fiber optic probe hydrophone system (*right panels*). In all panels, the acoustic source is a 788-kHz spherical shell transducer emitting a ten-cycle pulse. The transducer is driven using identical excitation conditions in (A) and (B), and then is also driven using identical higher-amplitude conditions in (C) and (D). No noise filtering is performed in any case, though the FOPH wave forms are averaged prior to recording [128-point averaging in (B), 16-point in (D)]. Though the FOPH wave forms are generally noisier, comparison of (C) and (D) demonstrates the ability of the FOPH system to more faithfully record the peak compressional pressures of shocked wave forms due to its enhanced bandwidth and smaller active aperture compared to those of the PVDF membrane hydrophone used.

emitting a 11-cycle pulse. These wave forms were measured with the FOPH system using different orientations of the fiber with respect to the beam axis. Figure 3(a) shows the wave form measured using the standard orientation with the plane of the fiber endface approximately normal to the beam axis (as depicted in Fig. 1), while Fig. 3(b) indicates the same wave form recorded with the fiber chuck positioned vertically in the tank, i.e. rotated 90° from the orientation used to generate Fig. 3(a) and now aligned such that the plane of the fiber endface is approximately parallel to the beam axis. Use of 100-μm diameter optical fiber avoids the need for spatial averaging corrections up to frequencies of ~7.5 MHz, which allows the FOPH to be considered practically omnidirectional at typical therapeutic ultrasound frequencies. Therefore, under these conditions, the fiber endface can be oriented arbitrarily with respect to the beam axis and can produce nearly identical wave forms with any alignment, as demonstrated in Fig. 3. It is worth noting that the slight ringing observed in the tail of the pressure signal in Fig. 3(a), which is attributed to reverberations from the fiber chuck housing, is reduced when the plane of the fiber endface is parallel to the beam axis in Fig. 3(b). However, in this configuration, the theoretical bandwidth of the FOPH device is significantly more limited than that predicted when the plane of the endface is normal to the beam axis.[6] Therefore, there is a tradeoff associated with aligning the endface plane to be parallel to the beam axis due to the reductions in both bandwidth and reverberation artifacts observed in this orientation. While parallel alignment of the fiber endface and the beam axis does appear to dampen the artifactual ringing observed in the signal, the restriction imposed on its bandwidth in this orientation implies that it should not be used in this manner to measure highly nonlinear wave forms with broad frequency content. For relatively low-amplitude wave forms with minimal shocking, however, arbitrary orientation of the optical fiber does produce equivalent wave forms, as Fig. 3 indicates.

C. Pressure amplitudes registered

The minimum detectable pressure registered using the FOPH system under reasonable averaging constraints (i.e., 256 point) was approximately 0.9 MPa. The largest compressional pressure registered was approximately 180 MPa and was largely limited by the bandwidth of the photodetector used in the FOPH system. Furthermore, peak rarefactional

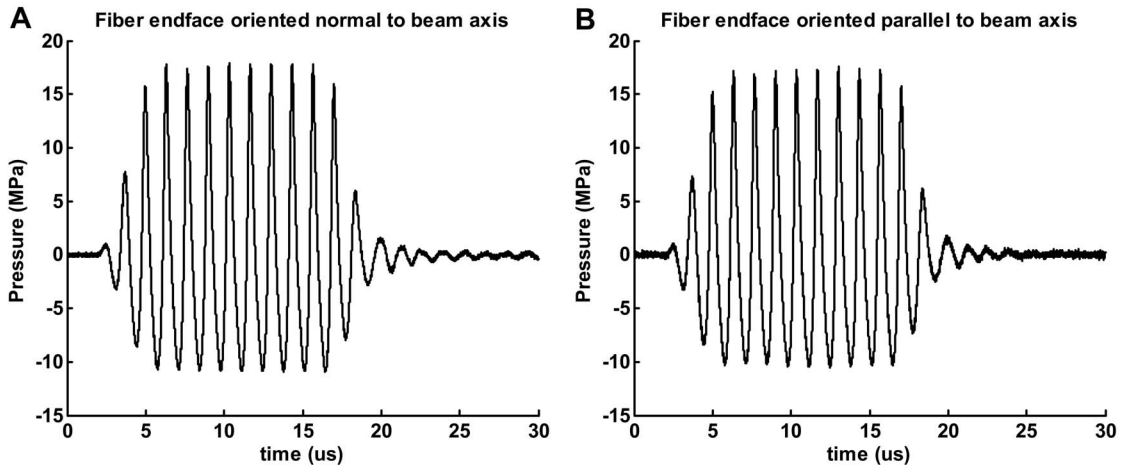


FIG. 3. Wave forms generated under identical excitation conditions measured with the FOPH system using different fiber orientations with respect to the beam axis. (A) Standard orientation with the fiber endface normal to the beam axis. (B) Alternate orientation with the fiber mounted vertically in the calibration tank so that the endface is parallel to the beam axis. In both panels, the acoustic source is a 750-kHz annular array emitting a 11-cycle pulse. For these wave forms of limited nonlinearity, very similar results are generated independent of the fiber orientation, with the exception that the parallel orientation depicted in (B) reduces ringing observed in the tail of the pressure signal that arises due to reverberations from the fiber chuck apparatus.

pressures of approximately 25 MPa could be registered without cavitation inception in the host water. This maximum rarefactional pressure measurement was conducted in well-degassed deionized water (10–20% gas saturation), but no special precautions were taken to assure the cleanliness of the water. If extreme care were used to ensure the purity of the host water, larger rarefactional pressures may be measurable without corruption of the hydrophone signal by cavitational activity. In the present authors' experience, the fiber was remarkably resistant to corruption by cavitation: For instance, the voltage signal was observed to become heavily distorted or to diminish completely during cavitation events, but frequently when cavitation activity ceased the voltage signal would reappear identically to its previous form without having to recleave the fiber tip.

V. DISCUSSION

A. Factors affecting measurement of pressure amplitudes

There are several sources of error inherent in the calibration methods for the FOPH device that affect both the precision of the transfer factors determined and the accuracy of resultant pressure amplitudes measured. During both the substitution and intrinsic calibration procedures described above, precision of the transfer factor measurement was limited by a standard deviation of approximately 5–6%. Factors contributing to this precision limit include those related to the stability of the dc reflectance level. Within a given calibration day, the dc reflectance was observed to vary within up to 2.6% of its mean from cleave to cleave, indicating that there may be some inherent variation in fiber cleaving and/or stripping techniques employed among users. However, the contribution of fiber cleaving and/or stripping irregularities to the transfer factor measurement error cannot be readily decoupled from that arising from instability in the optical power output of the laser diode, due to the fact that the dc reflectance was also observed to vary within up to 3.4% of its mean during measurements made with a single-cleaved fiber endface. Therefore, routine adjustment of the transfer factor should be performed by remeasuring the dc reflectance level each time the fiber is cleaved in order to account for cleaving and/or stripping variations that may arise. Furthermore, it may be advantageous to measure the static dc reflectance level immediately-before each wave form collection so that the transfer factor can be further adjusted to account for the minor temporal fluctuations in optical power output that were observed to occur.

In addition to the precision limit imposed by the above-mentioned variations in the dc reflectance, other factors also affect the accuracy of the pressures measured using the calibrated FOPH device. The most notable of these involves the use of the piezo-optic constant for water in the intrinsic determination of the transfer factor using Eq. (3) above, the value of which varies within $\sim$20% in the literature.[1,2,4,10,11] The present authors have opted to use the value provided by Phillips[1] because it is determined from first principles and appears to coincide with the experimental results described by Yadav *et al.*[10] However, to account for the range of values reported, other authors have considered an average piezo-optic constant of $\sim 1.5\times 10^{-4}$ MPa^{-1}.[4] In so doing, the intrinsic transfer factor of the present FOPH system is then calculated to be 0.70 ± 0.05 mV/MPa (-303.1 dB re 1 V/μPa; mean $\pm$ SD, $N=6$). The uncertainty associated with the reported values of the piezo-optic constant for water may provide impetus to implement alternate intrinsic calibration techniques as discussed, for example, by Staudenraus and Eisenmenger[2] and Wurster *et al.*[3] Furthermore, the uncertainty specified for the reference PVDF membrane hydrophone used during FOPH calibration via the substitution technique ($\pm$ up to 1.5 dB re 1 V/μPa for $f\leqslant 30$ MHz) poses an additional limit on the accuracy of the pressures registered by the FOPH device.

B. Suggested modifications to optical components

The FOPH system described herein serves as a device that functions acceptably without the use of filtering, external preamplifiers, or noise suppression techniques requiring additional photodetectors. However, due to the fact that the FOPH device does not employ any of these enhancements and the sensitivity is critically limited by the noise specifications of the optical components used, further optimization of this demonstration system will likely enable the development of transfer factors greater than the current value of ~0.8 mV/MPa. For instance, this transfer factor is ~20% lower than that reported by Huber *et al.*[6] (who did incorporate a wideband preamplifier and low-pass filtering to 10 MHz) and 2.5-fold lower than that of the commercially available FOPH device (RP Acoustics, Luetenbach, Germany). Furthermore, the transfer factor of the FOPH device presently described is approximately two times lower than that of a FOPH device recently reported by Mu *et al.*,[15] which is targeted at calibration of imaging transducers requiring up to 100-MHz bandwidth. The larger transfer factor reported by Mu *et al.*[15] is due at least in part to the use of an avalanche photodiode (APD) module with greatly enhanced responsivity compared to the *p-i-n* photodiode in use in the system presently discussed (~12 A/W @ 980 nm for the APD versus ~0.325 A/W @ 850 nm for the silicon *p-i-n* photodiode in the device reported by the present authors). However, the APD employed by Mu *et al.*[15] has a lower corner frequency (−3 dB) of 1 MHz, a feature that limits its usefulness with shock wave measurements or with therapeutic transducers that have center frequencies below this cutoff. Nevertheless, these comparisons stress the point that the photodetector characteristics critically affect both the expected transfer factor and the overall system bandwidth of the FOPH device and, as such, the detector should be chosen judiciously. Since the transfer factor of any FOPH device is directly proportional to both the incident optical power and the overall conversion gain of the photodetector (given by the product of its responsivity and transimpedance gain), the user could choose to maintain costs while increasing performance by incorporating a more sophisticated photodetector with higher conversion gain and lower noise floor at the expense of using a laser diode of lower optical power. In fact, this approach would be likely be significantly more advantageous in terms of boosting the expected signal level because the costs associated with increasing the photodetector conversion gain by a factor of 30 may be similar to those associated with incorporating a laser diode module with only a five-fold increase in optical power. In summary, though the basic FOPH system presented herein employs a high-power laser diode and a relatively inexpensive silicon *p-i-n* photodetector, a cost-benefit analysis would likely indicate that a lower-power optical source should be employed in order to facilitate the use of a photodetector with higher conversion gain.

C. Incorporation of FOPH transfer functions and fiber coatings

It should be noted that a wealth of information is available regarding further developments of the FOPH system that address a range of factors including, for example, the incorporation of the FOPH impulse response in corrections to the measured signal that are required due to the strongly varying system transfer function.[4,5,13,16] More accurate wave forms may be obtained by deconvolving the FOPH impulse response from the measured signal to rectify acoustic resonances and diffraction effects that develop due the sharp spatial discontinuity posed by the fiber edge. Other significant developments include the application of multilayer coatings to the fiber endface to improve sensitivity via interferometric techniques.[17–19] The use of dielectric or polymer films deposited on the fiber tip can increase the signal amplitude by factors of 30–300× compared to systems using bare sensors.[2,17,19] The fiber-coating modification may improve the FOPH system sensitivity sufficiently to allow its use with acoustic fields of lower amplitude than those that are practicable with the current device. Depositing such films onto the fiber tip to improve sensitivity may be of limited expediency with a FOPH device intended for use with high-amplitude acoustic fields, however, due to the fact that the user may expect to have to cleave and recoat the fiber tip repeatedly if corruption from cavitational activity is a frequent occurrence.

ACKNOWLEDGMENTS

The authors would like to thank Jochen Krücker for critical discussions regarding the operation of the hydrophone system and the initial determination of the laser diode module employed. This work has been supported by the National Institutes of Health Grants Nos. R01 DK42290 and 1R01 HL77629-01A1. J. Parsons is supported by the Whitaker Foundation.

[1] R. L. Phillips, "Proposed fiber-optic acoustical probe," Opt. Lett. **5**, 318–320 (1980).
[2] J. Staudenraus and W. Eisenmenger, "Fibre-optic probe hydrophone for ultrasonic and shock-wave measurements in water," Ultrasonics **31**, 267–272 (1993).
[3] C. Wurster, J. Staudenraus, and W. Eisenmenger, "The fiber optic probe hydrophone," in *Proceedings of the IEEE Ultrasonics Symposium*, 941–944 (Wiley-IEEE, Piscataway, 1994).
[4] J. Krücker, A. Eisenberg, M. Krix, R. Lotsch, M. Pessel, and H.-G. Trier, "Rigid piston approximation for computing the transfer function and angular response of a fiber-optic hydrophone," J. Acoust. Soc. Am. **107**, 1994–2003 (2000).
[5] Z. Q. Wang, R. Pecha, B. Gompf, and W. Eisenmenger, "Single bubble sonoluminescence: Investigations of the emitted pressure wave with a fiber optic probe hydrophone," Phys. Rev. E **59**, 1777–1780 (1999).
[6] P. Huber, J. Debus, P. Peschke, E. W. Hahn, and W. J. Lorenz, "*In vivo* detection of ultrasonically induced cavitation by a fibre-optic technique," Ultrasound Med. Biol. **20**, 811–825 (1994).
[7] P. J. Tait, "Report on some of the physical properties of fresh water and sea water," Phys. Chem. **2**, 1–76 (1888).
[8] J. H. Gladstone and T. P. Dale, "Researches on the refraction, dispersion, and sensitiveness of the liquids," Philos. Trans. R. Soc. London **153**, 317–343 (1864).
[9] L. Davison and R. A. Graham, "Shock compression of solids," Phys. Rep. **55**, 255–379 (1979).
[10] H. S. Yadav, D. S. Murty, S. N. Veerma, K. H. C. Sinha, B. M. Gupta, and D. Chand, "Measurement of refractive index of water under high dynamic pressures," J. Appl. Phys. **44**, 2197–2200 (1973).
[11] J. S. Rosen, "The refractive indices of alcohol, water, and their mixtures at high pressures," J. Opt. Soc. Am. **37**, 932–938 (1947).
[12] L. M. Barker and R. E. Hollenbach, "Shock wave studies of PMMA, fused

silica, and sapphire," J. Appl. Phys. **41**, 4208–4226 (1970).
[13]V. Wilkens and C. Koch, "Amplitude and phase calibration of hydrophones up to 70 MHz using broadband pulse excitation and an optical reference hydrophone," J. Acoust. Soc. Am. **115**, 2892–2903 (2004).
[14]G. M. Hale and M. R. Querry, "Optical constants of water in the 200-nm to 200-μm wavelength region," Appl. Opt. **12**, 555–563 (1973).
[15]C. Mu, S. Umchid, L. Bansal, P. A. Lewin, M. El-Sharif, and A. S. Daryoush, "Broadband fiber optic ultrasound hydrophone probe," in *Proceedings of the IEEE Lightwave Technologies in Instrumentation and Measurement Conference*, 149–152 (Wiley-IEEE, Piscataway, 2004).
[16]C. Koch, G. Ludwig, and W. Molkenstruck, "Calibration of an interferometric fiber tip sensor for ultrasound detection," Ultrasonics **35**, 297–303 (1997).
[17]A. J. Coleman, E. Draguioti, R. Tiptaf, N. Shotri, and J. E. Saunders, "Acoustic performance and clinical use of a fibreoptic hydrophone," Ultrasound Med. Biol. **24**, 143–151 (1998).
[18]W. Weise, V. Wilkens, and C. Koch, "Frequency response of fiber-optic multilayer hydrophones: Experimental investigation and finite element simulation," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **49**, 937–945 (2002).
[19]V. Wilkens and C. Koch, "Fiber-optic multilayer hydrophone for ultrasonic measurement," Ultrasonics **37**, 45–49 (1999).

Sound attenuation by small spheroidal particles due to visco-inertial coupling

Frank Babick[a] and Andreas Richter
Institut für Verfahrenstechnik und Umwelttechnik, TU Dresden, 01062 Dresden, Germany

(Received 27 May 2005; revised 1 January 2006; accepted 5 January 2006)

The sound attenuation at ultrasonic frequencies caused by small spheroidal particles in a fluid is examined with regard to the size parameters that determine the shape of the attenuation spectrum. Our investigations are based on a coupled phase model for spheroids with arbitrary orientation, thus facilitating the calculation of average attenuation for a given orientation distribution. Since the model just considers the visco-inertial coupling, its applicability is restricted to small solid particles with high density contrast. The calculated attenuation spectra of mono-sized, randomly aligned spheroidal particles are compared with the attenuation spectra of mono-sized spheres. When the latter approximate the former to a reasonable degree the size of the spheres is called attenuation equivalent diameter. It is shown that the concept of attenuation equivalent diameter can be applied only to slightly elongated prolates. Oblates and very stretched prolates yield considerably broader attenuation spectra than mono-sized spheres. While for oblates and for slightly elongated prolates the characteristic frequency of the attenuation spectrum is determined by the volume specific surface, no such attenuation determining size parameter could be identified for very stretched prolates. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2168427]

PACS number(s): 43.35.Bf, 43.20.Hq [YHB] Pages: 1441–1448

I. INTRODUCTION

During the past decade ultrasonic attenuation spectroscopy has become a well-accepted tool for the determination of particle size distributions of concentrated suspensions and emulsions.[1,2] The interpretation of acoustic attenuation spectra is based either on calibration data[3] or—more frequently—on analytical models.[4,5] For the sake of simplicity the commonly employed models assume particles of spherical shape. However, the majority of industrial suspensions consist of particles having shapes which considerably deviate from a sphere (e.g., clay particles, needlelike or plateletlike pigments, grinding products). Measured size distributions are then regarded as the diameter distributions of spheres with equivalent sound attenuation.[6] This interpretation actually requires that the attenuation spectrum of mono-sized nonspherical particles matches with the attenuation spectrum of spheres of a certain size. In fact, this assumption has never been verified. Besides, the concept of attenuation-equivalent spheres does not imply how the attenuation spectrum is related to other physical or geometrical properties of the particles. It therefore provides no answer to the question of whether the measured "particle size" is sensitive to those size parameters, which ultimately determine the product quality of the particle system.

The identification of the size parameters responsible for the sound attenuation of nonspherical particles may be achieved by theoretical considerations, which has been carried out in a number of studies. The most intensively examined case of nonspherical particles is presumably cylinders.[7–10] These papers actually examine the scattering of sound by cylindrical obstacles, from which sound attenuation can be derived. The scattering approach can principally cover the whole range of frequencies and particle sizes. However, to the author's knowledge, the scattering approach has not been used to quantify the attenuation of ellipsoids or even irregularly shaped particles.

Theoretical examinations of the sound attenuation by ellipsoidal particles have been conducted in the long wavelength regime ($ka \ll 1$), where the sound attenuation results from energy dissipation due to the viscous momentum transfer between the oscillating particles and the continuous phase. In suspensions with low density contrast there is an additional (sometimes dominating) contribution by the heat exchange between the particles and their vicinity. The theoretical treatment of sound attenuation in the long wavelength regime is simpler than in the general case, since scattering is negligible and the dissipated energy can be calculated from the interfacial momentum or energy flux and the corresponding driving force (e.g., the relative particle velocity).[11]

Regarding the sound attenuation due to the visco-inertial coupling between particles and fluid, this approach implies that shape effects have to be included via the oscillatory drag force. Expressions for this force are given by Lamb[12] and Happel and Brenner.[13] Ahuja and Hendee[14] use these expressions to describe the attenuation of prolate and oblate spheroids, with their symmetry axis aligned either perpendicular or parallel to the incident sound field. The authors do not provide calculations for spheroids with arbitrary orientation, but hint that in this case drag forces normal to the direction of sound propagation have to be considered.

Following the publication of Ahuja and Hendee, other authors have dealt with the acoustic attenuation by spheroids.[15–17] None of these papers lists detailed information on the models employed and the results obtained.

[a]Electronic mail: Frank.Babick@tu-dresden.de

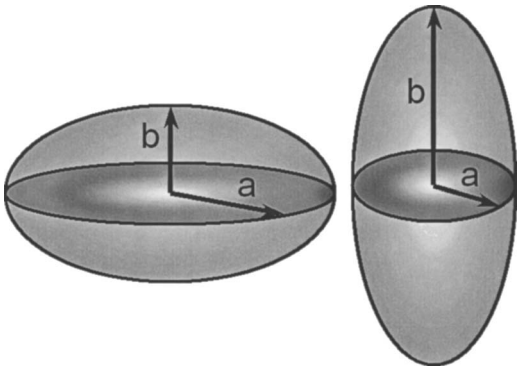


FIG. 1. Oblate and prolate spheroid; a denotes the radius of the equatorial plane, while b stands for the polar radius.

Pendse *et al.*[18] refer to the thesis of Han,[19] but omit the details of the employed theory. They point out that particle shape may be a distribution instead of a fixed value. Moreover, the particle shape may depend on particle size. The authors furthermore describe a technique to keep the particles aligned in the sensing zone, thus avoiding the problem of unknown particle orientation.

Beside the mentioned theoretical papers, there are few experimental works on the sound attenuation of nonspherical particles. Blue and McLeroy[20] present results for needle-shaped and platelet-shaped particles with similar distributions of settling velocity. They compare experimental spectra with the theoretical spectra of suspensions of spheres with the same settling velocity distribution. Even though their experimental results allow for ambiguous interpretation, there is a slight tendency that the agreement is better for platelet-like particles than for needlelike particles. Regardless, the comparison still lacks information on the geometrical size parameters that determine the sound attenuation.

A recent paper by Richards *et al.*[21] presents experimental results of very flat particles (platelets) and of more compact irregularly shaped particles. The measured attenuation spectra are compared with the theoretical predictions of Ahuja and Hendee.[14] The size distributions required for that are calculated from sedimentation experiments applying dynamic shape factors on the Stokes drag force. They found satisfying agreement for both particle systems. The results are not discussed in terms of the equivalent diameter concept.

The objectives of this paper are to present a model for the sound attenuation by spheroids due to the visco-inertial coupling for arbitrary orientations of the particles and to discuss the adequacy of the equivalent diameter concept when interpreting attenuation spectra of such nonspherical particles.

II. MODELING

A. Spheroids

Spheroids are ellipsoids having two axes of equal length (Fig. 1). Distinctions are made between oblate spheroids, whose polar axis is smaller than the equatorial axis ($b<a$), and prolate spheroids, which have a spindlelike shape ($b>a$). The deviation from spherical shape is characterized by the aspect ratio b/a or by the eccentricity $e(\sqrt{1-(b/a)^2}$ for oblates and $\sqrt{1-(a/b)^2}$ for prolates).

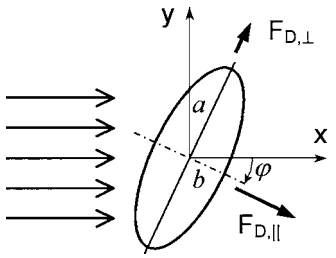


FIG. 2. Drag forces on a spheroid with its polar axis inclined by angle φ to the flow field.

B. Coupled phase model

In the long wavelength regime the attenuation of sound can be ascribed to the dissipative mechanisms of visco-inertial coupling and thermal coupling between particles and fluid as well as to the intrinsic attenuation within both phases. For solid particles having a high-density contrast ($\Delta\rho$) to the suspending medium only the visco-inertial coupling is significant. The sound propagation in such a suspension can be adequately modeled by so-called coupled phase models, which have been extensively described.[5,22–24] The idea behind this theoretical approach is that the suspension can be regarded as two separate phases (particles and fluid), which are coupled by momentum transfer. Exerting a sound field to such a system means that the equations of state, continuity, and momentum conservation can be fulfilled simultaneously only for the characteristic wave number of the suspension. The approach is strictly limited to small wave numbers ($ka \ll 1$).

For spheroidal particles arbitrarily aligned to an incident plane wave the equations of momentum transfer assume the following form (cf. Fig. 2):

$$\rho_P \frac{\partial v_{P,x}}{\partial t} = -\frac{\partial p}{\partial x} + f_{D,\perp} \sin\varphi + f_{D,\parallel} \cos\varphi, \tag{1}$$

$$\rho_P \frac{\partial v_{P,y}}{\partial t} = f_{D,\perp} \cos\varphi - f_{D,\parallel} \sin\varphi, \tag{2}$$

$$(1-\phi)\rho_F \frac{\partial v_{F,x}}{\partial t} = -(1-\phi)\frac{\partial p}{\partial x} - \phi \cdot f_{D,\perp} \sin\varphi - \phi f_{D,\parallel} \cos\varphi, \tag{3}$$

$$(1-\phi)\rho_F = \frac{\partial v_{F,y}}{\partial t} = -\phi \cdot f_{D,\perp} \cos\varphi + \phi f_{D,\parallel} \sin\varphi, \tag{4}$$

where p is the excess pressure, $v_{x/y}$ are the Cartesian velocity components, ρ is the density, ϕ is the volume fraction of the particles, φ is the angle between the wave vector $\mathbf{k}$ and the polar axis, and $f_{D,\parallel/\perp}$ is the volume specific drag force ($f_D = F_D/V_P$) on a particle that is aligned parallel or perpendicular to the polar axis. The indices F and P denote the fluid and particle phase, respectively.

The drag forces are determined by the relative velocity between fluid and particle, thus

$$f_{D,\perp} = -C_{D,\perp}((\mathbf{v}_P - \mathbf{v}_F)_x \sin\varphi + (\mathbf{v}_P - \mathbf{v}_F)_y \cos\varphi), \quad (5)$$

$$f_{D,\parallel} = -C_{D,\parallel}((\mathbf{v}_P - \mathbf{v}_F)_x \cos\varphi - (\mathbf{v}_P - \mathbf{v}_F)_y \sin\varphi). \quad (6)$$

The drag coefficient C_D can be split into a dissipative and an inertial part,[14]

$$C_D = \omega\rho_F(s - i\tau), \quad (7)$$

where ω denotes the angular frequency and i the imaginary number. For isolated spheroids the dissipative drag coefficient s is given by

$$s = \frac{9}{4}\frac{\delta_{vis}}{b}K^2\left(1 + \frac{1}{K}\frac{\delta_{vis}}{a}\right) \quad (8)$$

and the inertial drag coefficient τ assumes

$$\tau = L + \frac{9}{4}\frac{\delta_{vis}}{b}K^2, \quad (9)$$

where δ_{vis} is the viscous penetration depth ($\delta_{vis} = \sqrt{2\eta_F/\rho_F\omega}$, η_F is the fluid viscosity). The parameter K is the dynamic shape factor and L is the coefficient for the induced mass.

Expressions for these parameters can be drawn from the paper by Ahuja and Hendee.[14] For the purpose of convenience they are given in the Appendix .

The four equations of motion [(1)–(4)] do not suffice to describe the system completely. The missing equation is obtained by combining the material equation for the relationship between density and pressure:

$$\frac{d\rho_j}{\rho_j} = -\kappa_{ad,j}dp \quad (10)$$

with the continuum balance

$$\frac{\partial\rho_j}{\partial t} = \mathrm{div}(\rho_j\mathbf{v}_j) = \rho_j \cdot \frac{\partial v_{j,x}}{\partial x} \quad (11)$$

to

$$-\langle\kappa_{ad}\rangle\frac{\partial p}{\partial t} = \phi\frac{\partial v_{P,x}}{\partial x} + (1-\phi)\frac{\partial v_{F,x}}{\partial x}, \quad (12)$$

where κ_{ad} denotes the adiabatic compressibility and $\langle\cdots\rangle$ is a volume weighted average. Note that the y components of the divergence operator vanish, since a plane wave is assumed.

Assuming harmonic oscillations and wave propagation in the x direction [$p \propto \exp(-i(\omega t - kx))$], the following system of equations is achieved:

$$0 = -ikp - (A - i\omega\rho_P)v_{P,x} - Dv_{P,y} + Av_{F,x} + Dv_{F,y}, \quad (13)$$

$$0 = -Dv_{P,x} - (B - i\omega\rho_P)v_{P,y} + Dv_{F,x} + Bv_{F,y}, \quad (14)$$

$$0 = -ik(1-\phi)p + \phi Av_{P,x} + \phi Dv_{P,y} - (\phi A - i\omega(1-\phi)\rho_F)v_{F,x} - \phi Dv_{F,y}, \quad (15)$$

$$0 = \phi Dv_{P,x} + \phi Bv_{P,y} - \phi Dv_{F,x} - (\phi B - i\omega(1-\phi)\rho_F)v_{F,y}, \quad (16)$$

$$0 = -i\omega\langle\kappa_{ad}\rangle p + ik\phi v_{P,x} + ik(1-\phi)v_{F,x}, \quad (17)$$

where

$$A = C_{D,\perp}\sin^2\varphi + C_{D,\parallel}\cos^2\varphi, \quad (18)$$

$$B = C_{D,\perp}\cos^2\varphi + C_{D,\parallel}\sin^2\varphi, \quad (19)$$

$$D = (C_{D,\perp} - C_{D,\parallel})\sin\varphi\cos\varphi. \quad (20)$$

Note that in contrast to spheres or spheroids with one principal axis aligned to the incident sound field there is a drag force component and hence velocity component normal to the direction of sound propagation (y axis, see Fig. 2).

The system yields a nontrivial solution only for a vanishing determinant of the coefficient matrix. The wave number k corresponding to that case is considered to be the wave number k_s of the suspension. By rearranging the equation system and calculating the corresponding determinant the wave number k_s can be resolved:

$$\frac{k_s^2}{\omega^2} = \langle\kappa_{ad}\rangle\langle\rho\rangle\frac{(C_{D,\parallel} - i\omega(1-\phi)(\rho_F\rho_P/\langle\rho\rangle))(C_{D,\perp} - i\omega(1-\phi)(\rho_F\rho_P/\langle\rho\rangle))}{(A - i\omega(1-\phi)\rho_F\rho_P\langle 1/\rho\rangle)(B - i\omega(1-\phi)(\rho_F\rho_P/\langle\rho\rangle)) - D^2}. \quad (21)$$

For spheroids aligned parallel and perpendicular to the incident field, this solution reduces to

$$\frac{k_s^2}{\omega^2} = \langle\kappa_{ad}\rangle\langle\rho\rangle\frac{\frac{C_D}{\omega} - i(1-\phi)(\rho_F\rho_P/\langle\rho\rangle)}{\frac{C_D}{\omega} - i(1-\phi)\rho_F\rho_P\langle 1/\rho\rangle}, \quad (22)$$

where $C_D = C_{D,\parallel}$ for parallel and $C_D = C_{D,\perp}$ for perpendicular orientation. For spherical particles ($C_D = C_{D,\perp} = C_{D,\parallel}$) the mathematical structure of this equation is formally identical to the solutions of previous workers,[5,23,24] even though the drag coefficients used by these authors are defined differently. From Eq. (22) the attenuation coefficient is derived as

$$\alpha = \frac{\phi}{2}(1-\phi)^2 c_s\langle\kappa_{ad}\rangle\frac{(\rho_P/\rho_F - 1)^2\omega\rho_F s}{(\rho_P\langle 1/\rho\rangle(1-\phi) + \tau)^2 + s^2}, \quad (23)$$

where c_s represents the sound speed of the suspension. This expression converges to Eq. (15) of Ahuja and Hendee[14] for

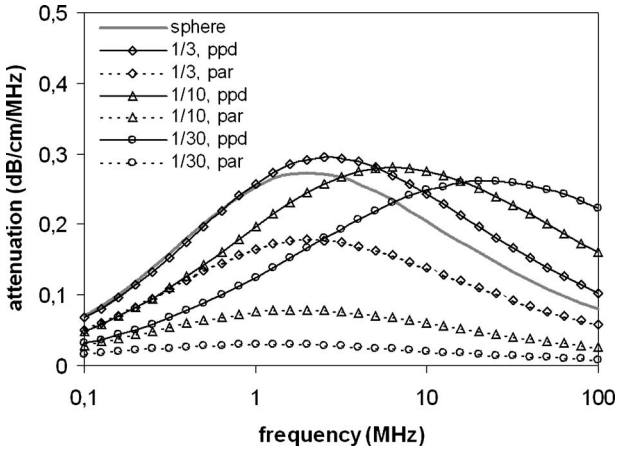


FIG. 3. Attenuation spectra of oblate particles with their polar axis aligned either perpendicular (solid lines) or parallel (dotted lines) to the incident sound field; the aspect ratio varies from 1/1 (sphere) to 1/30 and the equatorial radius from 0.5 μm ($ka_{10\ \mathrm{MHz}}$=0.02) to 1.55 μm ($ka_{10\ \mathrm{MHz}}$=0.065) while the volume is kept constant; the particle volume fraction is 0.01.

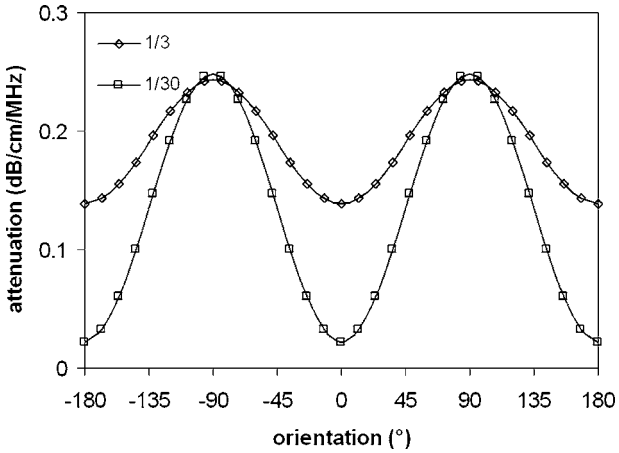


FIG. 5. Sound attenuation of oblate spheroids versus the angle of incidence; the sound frequency is 10 MHz, the aspect ratio varies from 1/3 to 1/30, and the equatorial radius from 0.72 μm (ka=0.03) to 1.55 μm (ka=0.065) while the volume is kept constant; the particle volume fraction is 0.01.

very dilute suspensions ($\phi \to 0$), when the effective properties $\langle \cdots \rangle$ and c_s assume the fluid values and $(1-\phi)$ equals 1. Note that Eq. (23) a priori accounts for the fact that the momentum of the fluid phase is altered by the momentum exchange. However, the influence of the particle concentration on the hydrodynamic conditions has to be included via modified drag coefficients s and τ.

III. RESULTS

Figures 3 and 4 show the attenuation spectra of suspensions of mono-sized oblates and prolates, respectively. While the aspect ratio is varied, the particle volume is kept constant to allow for comparability. The particles are considered to be either aligned perpendicular or parallel to the incident field. Calculations were conducted with water as the continuous phase (25 °C, ρ=997 kg/m^3, η=0.000 89 Pa s) and silica as the disperse phase (ρ=2200 kg/m^3). The results are presented as attenuation coefficient divided by frequency for frequency values from 0.1 to 100 MHz. This range is relevant for most commercial ultrasonic spectrometers used in particle sizing. The diameter of the volume-equivalent sphere (1 μm) was chosen in such a way that on the one side the attenuation maxima lie within the frequency range and that on the other side the long wavelength condition is fulfilled.

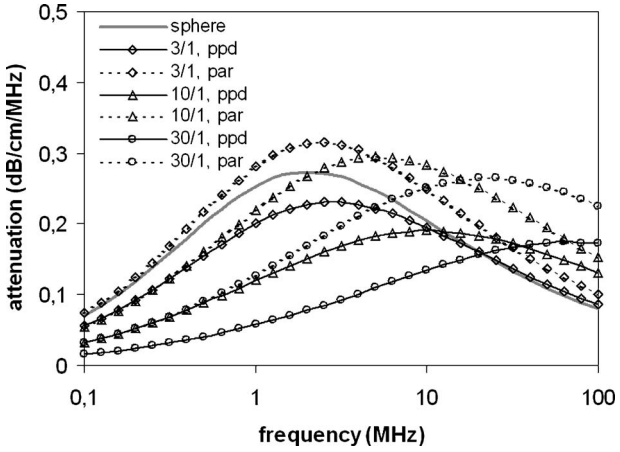


FIG. 4. Attenuation spectra of prolate particles with their polar axis aligned either perpendicular (solid lines) or parallel (dotted lines) to the incident sound field; the aspect ratio varies from 1/1 (sphere) to 30/1 and the polar radius from 0.5 μm ($kb_{10\ \mathrm{MHz}}$=0.02) to 4.8 μm ($kb_{10\ \mathrm{MHz}}$=0.20) while the volume is kept constant; the particle volume fraction is 0.01.

The results are in qualitative agreement with those published by Ahuja and Hendee,[14] showing a significant influence of orientation and aspect ratio on the shape and location of the attenuation spectra.

For oblate spheroids aligned perpendicular to the exciting field the attenuation values have the same order of magnitude as the volume-equivalent sphere, whereas oblates with a parallel orientation show considerably lower attenuation values. Besides, there is an increasing shift of the attenuation maximum to higher frequencies with increasing deviation from spherical shape when the oblates are perpendicularly oriented, but there is apparently no such effect for parallel aligned ones. In comparison, for prolate spheroids the frequency shift of the attenuation maximum can be observed for both principal orientations. Besides, perpendicular orientation results in lower attenuation values than parallel orientation, which is different from the oblates.

The distinct orientation effect in the sound attenuation corresponds to the orientation dependency of the relative particle velocity: For example in the case of oblate spheroids a broadside orientation of the particles results in a minimum relative velocity and in minimum attenuation, while for an edgewise orientation both parameters take on maximum values.[14]

Figures 3 and 4 can provide useful estimates for the attenuation of suspensions with randomly oriented particles (or with a certain orientation distribution), since they mark the limits of attenuation expectations for such systems. However, for spheroids with high eccentricity the difference between perpendicular and parallel orientation may become rather large (in particular for disc-shaped particles). It is therefore inevitable to know the attenuation dependency on orientation. For selected spheroids the attenuation values are

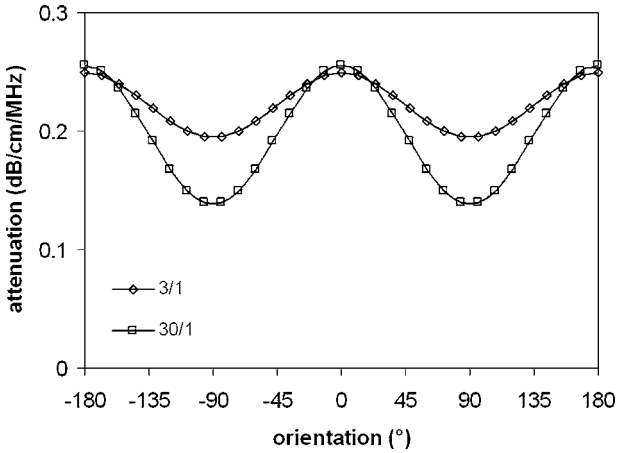


FIG. 6. Sound attenuation of prolate spheroids versus the angle of incidence; the sound frequency is 10 MHz, the aspect ratio varies from 3/1 to 30/1 and the polar radius from 1.0 μm (kb=0.044) to 4.8 μm (kb=0.20) while the volume is kept constant; the particle volume fraction is 0.01.

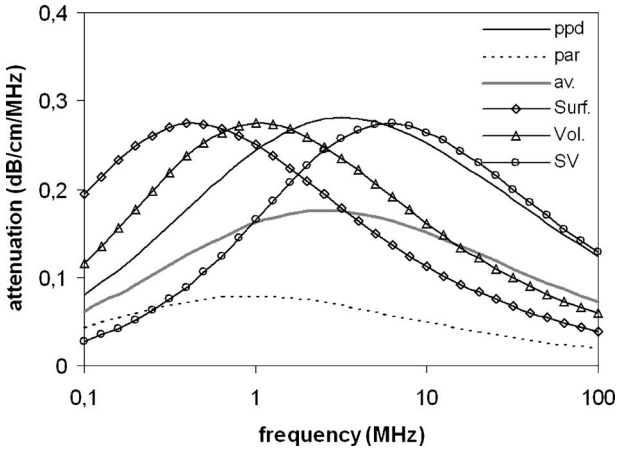


FIG. 7. Attenuation spectra of perpendicular, parallel, and arbitrarily aligned oblate particles with an aspect ratio of 1/10 and the corresponding attenuation spectra of spheres equivalent with respect to surface, volume, and volume-specific surface; the equatorial radius a of the oblate is 1.5 μm, the equivalent sphere radii are 1.1, 0.70, and 0.29 μm, respectively; the particle volume fraction is 0.01.

shown as a function of the angle of orientation in Figs. 5 and 6. The calculation parameters are as used for Figs. 3 and 4.

According to these graphs, the orientation dependency can be described as a harmonic function for both types of spheroids and all aspect ratios. That means that in the case of randomly orientated particles the mean attenuation can be calculated as the average of the attenuation values for perpendicular and parallel orientation or as the attenuation value at an orientation of 45°.

Real angle distributions are affected by the flow regime and gravity and may not be regarded as uniform. If these distribution functions are known, the harmonic function of Figs. 5 and 6 have to be weighted with them. In most practical applications the assumption of a uniform angle distribution should be a good approximation.

IV. DISCUSSION

Based on the derived coupled phase model and the approximation of uniform angle distribution the attenuation by spheroids shall now be compared with the attenuation by spheres. If the spectrum of the spheroids can be approximated by a spectrum of spheres, the size of the spheres can be regarded as the "acoustic attenuation equivalent diameter." This concept is quite common in particle sizing.[6] A next question arising from this comparison is how such an attenuation equivalent diameter relates to the equivalent diameters based on geometrical particle properties.

For this purpose the attenuation spectra of oblate and prolate spheroids (aspect ratios: 1/10 and 10/1, respectively) are compared with the attenuation spectra of their volume equivalent, surface equivalent, and volume specific surface ($S_V = S_P / V_P$) equivalent spheres. The material system and frequency range are as used above (Figs. 3 and 4). The size of the spheroids was chosen such that the attenuation maxima are within the frequency range and the long wavelength condition is obeyed.

For oblate spheroids Fig. 7 shows that the deviation of the orientation averaged attenuation spectra from any spectrum of mono-sized spheres is so large that it is not permissible to speak of attenuation equivalent spheres. In comparison this deviation is less pronounced for the prolate spheroids (Fig. 8). It appears that the attenuation spectrum of the examined prolates can be roughly approximated by the spectrum of spheres with the same volume-specific surface. However, the spectra in Fig. 7 and 8 were calculated for fixed aspect ratios. It can be expected that this parameter affects the extent to which a spheroid spectrum agrees with the one of spheres with the same specific surface area.

In Figs. 9 and 10 the attenuation spectra of randomly aligned spheroids with equal volume specific surface but varying aspect ratio are presented. The parameters for the calculation with regard to the material system and the diameter of the equivalent sphere are the same as for Figs. 3 and 4. The aspect ratio is varied from 1/1 (sphere) to 300/1 for the prolate spheroids and from 1/1 to 1/100 for the oblate spheroids. Additionally the limiting case of an ellipsoidal

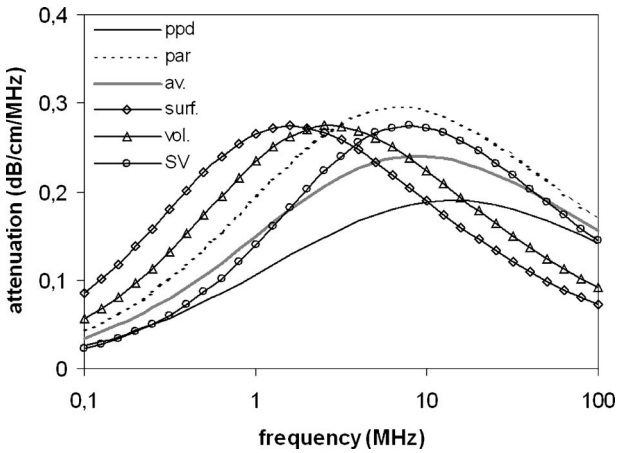


FIG. 8. Attenuation spectra of perpendicular, parallel, and arbitrarily aligned prolate particles with an aspect ratio of 10/1 and the corresponding attenuation spectra of spheres equivalent with respect to surface, volume, and volume-specific surface; the polar radius b of the prolate is 1.0 μm, the equivalent sphere radii are 0.56, 0.42, and 0.25 μm, respectively; the particle volume fraction is 0.01.

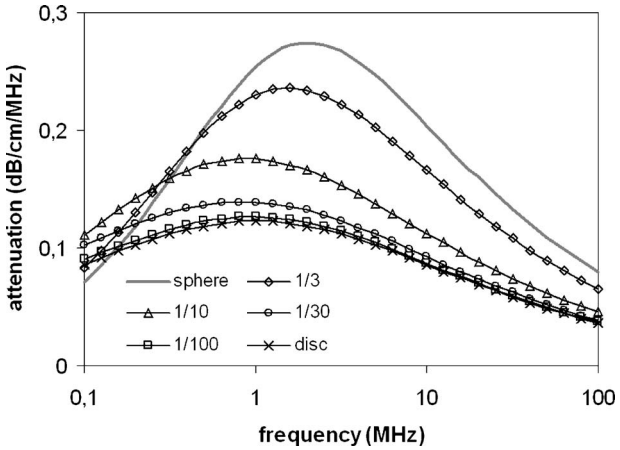


FIG. 9. Attenuation spectra of arbitrarily aligned oblate particles and ellipsoidal discs; the aspect ratio varies from 1/1 (sphere) to 1/100 and the equatorial radius from 0.5 μm ($ka_{10\ \mathrm{MHz}}$=0.02) to 25 μm ($ka_{10\ \mathrm{MHz}}$=1.1) while the volume-specific surface is kept constant; for the ellipsoidal discs the equatorial radius is undefined, their thickness is 0.5 μm; the particle volume fraction is 0.01.

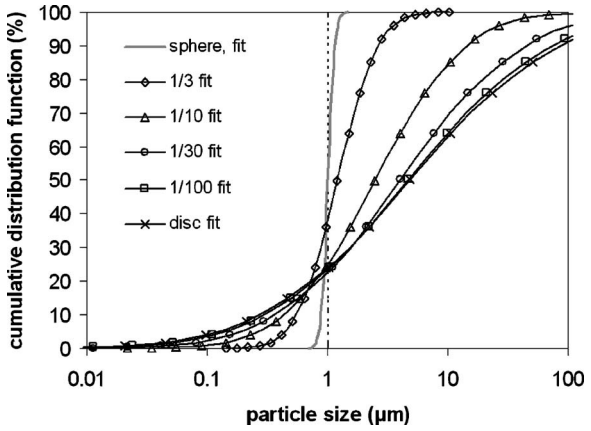


FIG. 11. Particle size distributions calculated from the spectra of Fig. 9 under the assumption of spherical particles and a log normal distribution; the dotted line indicate the diameter of a sphere with same volume-specific surface as the oblates from which the spectra in Fig. 9 were generated.

disc with infinite polar radius but defined thickness is given in the oblate diagram (Fig. 9). Note that at aspect ratios below 1/30 and above 30/1 the long wavelength condition is no longer fulfilled in the upper frequency range for the used spheroid dimensions. Then scattering would dominate the sound attenuation and the coupled phase model would no longer be an adequate representation of physics.

Figure 9 makes it clear that the deviation of the oblate spectrum from a sphere spectrum is quite significant (Fig. 9) even for low eccentricity (i.e., aspect ratios higher than 1/10). In particular the maximum attenuation falls well below the sphere value. Moreover, the maximum shifts slightly to lower frequencies. It converges, however, to a defined value, which is half the characteristic frequency of the sphere spectrum. At low aspect ratios (i.e., 1/10 and less) the volume-specific surface of oblates is solely determined by the polar radius b or the thickness of the oblate ($S_V \sim 1.5/b$). Consequently, we can state that the attenuation of oblates is chiefly controlled by their volume-specific surface, but that the spectra cannot be approximated by that of mono-sized spheres.

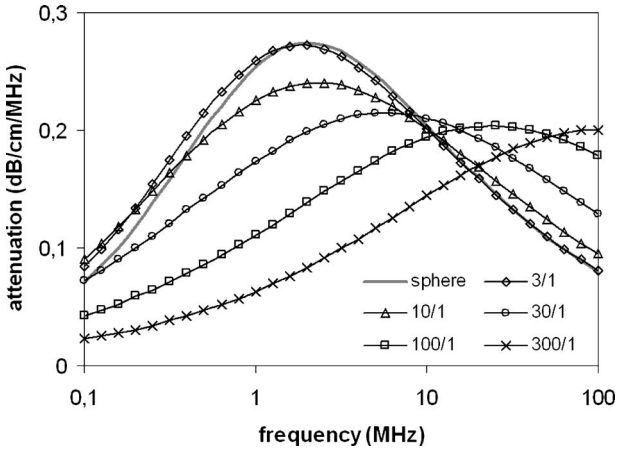


FIG. 10. Attenuation spectra of arbitrarily aligned prolate particles; the aspect ratio varies from 1/1 (sphere) to 300/1 and the polar radius from 0.5 μm ($kb_{10\ \mathrm{MHz}}$=0.02) to 118 μm ($kb_{10\ \mathrm{MHz}}$=4.9) while the volume-specific surface is kept constant; the particle volume fraction is 0.01.

A different picture is observed for prolate spheroids: For low eccentricities, i.e., for aspect ratios not larger than 3/1, the spectrum of mono-sized spheres with the same volume-specific surface as the prolates provides a good estimate for the attenuation of the spheroids (Fig. 10). However, for an aspect ratio of about 10/1 the estimate is only rough and at higher aspect ratios, i.e., more needlelike particles or fibers, the spheroid spectra not only show a lower maximum attenuation and a broader shape but also a shift of the attenuation maximum to higher frequencies. This contradicts the assumption that the volume-specific surface is the attenuation determining size parameter. For spheres of different size a mastercurve of all attenuation spectra can be obtained by scaling the x axis with a^2f, where a is the sphere radius and f is the sound frequency. According to this, the attenuation determining size parameter of very stretched prolate spheroids should be smaller than the radius (or diameter) of the sphere with equivalent volume-specific surface. Since the volume-specific surface of such spheroids can be approximated from the equatorial radius a by $3\pi/4a$, our results imply that this size parameter should be considerably smaller than the equatorial radius. Such a parameter will probably not possess a geometrical meaning.

In general, the spheroid spectra are smoother, i.e., flatter and broader, than the sphere spectra. Consequently, their interpretation with a model for spherical particles will result in an overestimation of the distribution width. This is illustrated in Figs. 11 and 12, which show the size distribution calculated from the attenuation spectra in Figs. 9 and 10, respectively. For the calculation, spherical particles with a log normal size distribution are assumed. The parameter of the distribution, the volume-weighted median size $x_{50,3}$, and the logarithmic standard deviation $\sigma_{\ln}$ are determined by fitting the mono-sized spheroid spectrum with a spectrum of polydisperse spheres. The fitting algorithm employed a modified gradient method with an allowable parameter range of 0.01 to 10 μm for the median size and 0.01 to 3 for the logarith-

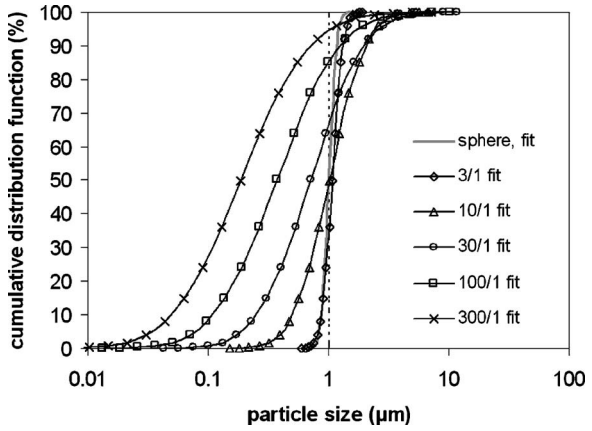


FIG. 12. Particle size distributions calculated from the spectra of Fig. 10 under the assumption of spherical particles and a log normal distribution; the dotted line indicates the diameter of a sphere with same volume-specific surface as the prolates from which the spectra in Fig. 10 were generated.

mic standard deviation. The algorithm has been proven to perform very well over a wide range of parameters, but has difficulties resolving very narrow distributions.[25] That is why the analysis yields a certain distribution width for the sphere spectrum ($\sigma_{\ln}=0.11$).

In general the calculated size distributions correspond with the attenuation spectra: For oblates—even with low eccentricity—the best fit is obtained for relatively large distribution widths ($\sigma_{\ln}=0.6$ for an aspect ratio of 1/3 and $\sigma_{\ln}=2.2$ for the ellipsoidal disc). According to the explanations given above, the median size increases with decreasing aspect ratio. In general, the fit is rather poor (Fig. 13).

In comparison, the attenuation spectra of the prolate spheroids are fitted rather well even for high aspect ratios (Fig. 13). However, only for the lowest aspect ratios in the calculation (3/1) does the determined size distribution indicate a mono-sized particle system ($\sigma_{\ln}=0.17$). For the aspect ratios 10/1 and 300/1 the logarithmic standard deviation $\sigma_{\ln}$ assumes values of 0.55 and 1.0, respectively. As already explained above, the determined sphere diameters are considerably smaller than the diameter of the sphere with the same volume-specific surface as the prolate spheroids ($x_{50,3}=0.19$ μm for aspect ratio of 300/1 as compared to a diameter of 1 μm).

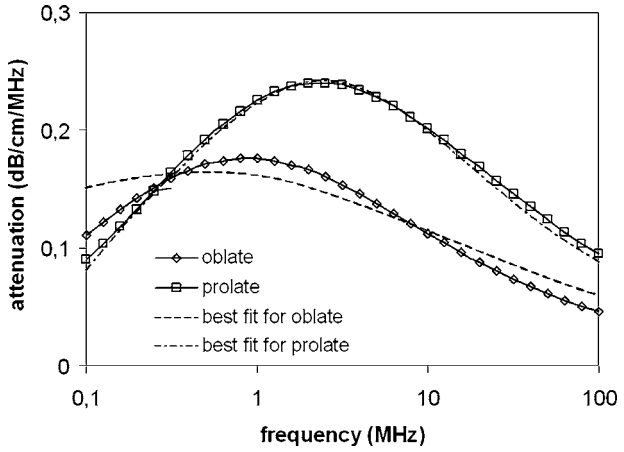


FIG. 13. Attenuation spectra of arbitrarily aligned, mono-sized prolate and oblate particles with aspect ratios of 10/1 and 1/10, respectively (diameter of spheres equivalent with respect to volume-specific surface is 1 μm) and best fit of this spectra when spherical particles and a log normal size distribution are assumed.

V. CONCLUSION

The primary objective of this paper is to evaluate if the concept of equivalent diameter—often used in particle sizing—can be applied to the interpretation of attenuation spectra of nonspherical particles. The study was conducted on the grounds of theoretical attenuation spectra of spheroids. First, a theoretical model is presented that allows for arbitrary particle orientations. This model predicts a harmonic dependency of attenuation on orientation angle. Therefore, the average attenuation of randomly oriented spheroids can be calculated as the mean of a perpendicular and a parallel orientation.

Applying the model to oblates and prolates with different aspect ratio reveals that the two spheroid types distinguish from each other with regard to the ultrasonic behaviour. For oblates the attenuation is mainly determined by the volume-specific surface. But the spectra deviate considerably in magnitude and shape from the attenuation spectra of spheres even for rather low eccentricity. When size distributions are calculated from the spectra of mono-sized oblates under the assumption of spherical particles, one obtains relatively broad size distributions. Prolate spheroids that are not very stretched (i.e., with aspect ratios below 10/1) attenuate sound quite similar to a system of spheres with the same volume-specific surface. Size distributions calculated from the spectra of such spheroids reproduce the distribution width reasonably well. However, elongated prolates with aspect ratios above 10/1 generate spectra that can be approximated by sphere spectra only for large distribution widths. Moreover, the attenuation controlling size parameter can no longer be approximated by the volume-specific surface, but is considerably smaller than the equatorial radius of the prolates. Hence, regarding all examined cases the concept of an attenuation equivalent sphere is applicable only for not too stretched prolate spheroids (aspect ratio $<10/1$).

The applicability of the results given above is restricted to the case of spheroid particles, though this shape may often be a good approximation of real particle shapes. However, attenuation may not only depend on macroshape but also on meso- or microshape parameters (like concavities and surface roughness, respectively). Qian[26] showed that for irregularly shaped sand particles sound attenuation is related to the fractal dimension. A further restriction results from neglecting scattering and thermal coupling. Hence the model is applicable only to watery suspensions of fine solid particles with high density contrast ($ka \ll 1$, $\Delta\rho > 500$ kg/m^3).

In general, there is still a lack of systematic experimental tests regarding the influence of particle shape on sound attenuation. Such data would allow for an evaluation of the applicability of the model given in this paper and the conclusions drawn from its predictions.

APPENDIX: Dynamic shape factors and coefficients of induced mass for spheroid particles

1. Dynamic shape factor

The dynamic shape factors K of spheroids were derived by Happel and Brenner[13] for flow parallel and perpendicular to the polar axis, respectively. For oblates with $h=b/a=\sqrt{1-e^2}<1$ they are

$$K_{\parallel}=\frac{\frac{4}{3}(1-h^2)}{[(1-2h^2)/(\sqrt{1-h^2})]\arctan[(\sqrt{1-h^2})/h]+h}, \tag{A1}$$

$$K_{\perp}=\frac{\frac{8}{3}(1-h^2)}{[(3-2h^2)/\sqrt{1-h^2}]\arctan(\sqrt{1-h^2}/h)-h}. \tag{A2}$$

For prolate spheroids the aspect ratio is $h=b/a=1/\sqrt{1-e^2}>1$ and the shape factors are

$$K_{\parallel}=\frac{\frac{4}{3}(h^2-1)}{[(2h^2-1)/\sqrt{h^2-1}]\ln(h+\sqrt{h^2-1})-h}, \tag{A3}$$

$$K_{\perp}=\frac{\frac{8}{3}(h^2-1)}{[(2h^2-3)/\sqrt{h^2-1}]\ln(h+\sqrt{h^2-1})+h}. \tag{A4}$$

Coefficient of induced mass

Expressions for the coefficient L for the induced mass of spheroids were given by Lamb.[12] For oblate spheroids they are

$$L_{\parallel}=\frac{e-\sqrt{1-e^2}\arcsin(e)}{\sqrt{1-e^2}\arcsin(e)-(1-e^2)e}, \tag{A5}$$

$$L_{\perp}=\frac{(1-e^2)e-\sqrt{1-e^2}\arcsin(e)}{\sqrt{1-e^2}\arcsin(e)-(1+e^2)e}, \tag{A6}$$

and for prolate spheroid

$$L_{\parallel}=\frac{1-(1/2e)\ln[(1+e)/(1-e)]}{(1/2e)\ln[(1+e)/(1-e)]-[1/(1-e^2)]}, \tag{A7}$$

$$L_{\perp}=\frac{1-[(1-e^2)/2e]\ln[(1+e)/(1-e)]}{[(1-e^2)/2e]\ln[(1+e)/(1-e)]+2e^2-1}. \tag{A8}$$

Note that the validity of these equations for both parameters, the induced mass coefficient and the dynamic shape factor, is restricted to dilute suspensions.

[1] D. J. McClements, "Principles of Ultrasonic Droplet Size Determination in Emulsions," Langmuir **12**, 3454–3461 (1996).
[2] A. S. Dukhin and J. P. Goetz, "Acoustic and electroacoustic spectroscopy for characterizing concentrated dispersions and emulsions," Adv. Colloid Interface Sci. **92**, 73–132 (2001).
[3] M. Li, P. Mougin, and D. Wilkinson, "On-line Crystallisation Monitoring by Ultrasonic Spectroscopy" on the CD-ROM: *Nuremberg, 16.-18.03.2004, Proceedings, PARTEC 2004, International Congress for Particle Technology*, available from NürnbergMesse GmbH, Messezentrum, 90471 Nürnberg, Germany.
[4] J. R. Allegra and S. A. Hawley, "Attenuation of sound in suspensions and emulsions: Theory and Experiments," J. Acoust. Soc. Am. **51**, 1545–1564 (1972).
[5] A. S. Dukhin and P. J. Goetz, "Acoustic Spectroscopy for Concentrated Polydisperse Colloids with High Density Contrast," Langmuir **12**, 4987–4997 (1996).
[6] K. Leschonski, "Particle Size Analysis and Characterization of Classification Processes," in *Ullmann's Encyclopaedia of Industrial Chemistry*, 6th ed., Vol. 25 edited by J. E. Bailey (Wiley-VCH-Verlag, Weinheim, 2002), pp. 183–218.
[7] C. J. T. Sewell, "The Extinction of Sound in a Viscous Atmosphere by Small Obstacles of Cylindrical and Spherical Form," Philos. Trans. R. Soc. London, Ser. A **210**, 239–270 (1910).
[8] K. Attenborough and L. A. Walker, "Sound dissipation by a small cylindrical obstacle," J. Acoust. Soc. Am. **51**(1, Pt. 2), 192–196 (1972).
[9] C. C. Habeger, "The attenuation of ultrasound in dilute polymeric fiber suspensions," J. Acoust. Soc. Am. **72**, 870–878 (1983).
[10] W. H. Lin and A. C. Raptis, "Thermoviscous effects of acoustic scattering by thermoelastic solid cylinders and spheres," J. Acoust. Soc. Am. **74**, 1542–1554 (1983).
[11] R. J. Urick, "The absorption of sound in suspensions of irregular particles," J. Acoust. Soc. Am. **20**, 283–289 (1948).
[12] H. Lamb, *Hydrodynamics*, 6th ed. (Cambridge U. P., Cambridge, 1959).
[13] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics* (Nijhoff, Dordrecht, 1973).
[14] A. S. Ahuja and W. R. Hendee, "Effects of particle shape and orientation on propagation of sound in suspensions," J. Acoust. Soc. Am. **63**, 1074–1080 (1978).
[15] K. Attenborough, Q. Wang, and S. Woodhead, "New models and measurements concerning sound attenuation in concentrated airborne suspensions," J. Acoust. Soc. Am. **107**(5, Pt. 2), 2846 (2000).
[16] A. S. Schaafsma, "Attenuation spectroscopy of suspended sediment suspensions," J. Acoust. Soc. Am. **92**, 2308 (1992).
[17] W. Han and H. H. Pendse, "Effect of particle shape on viscous attenuation of ultrasound in concentrated clay suspensions," J. Acoust. Soc. Am. **97**(5, Pt. 2), 3373 (1995).
[18] H. P. Pendse, T. C. Bliss, and W. Han, "Particle Shape Effects and Active Ultrasonic Spectroscopy," in *Handbook on Ultrasonic and Dielectric Characterization Techniques for Suspended Particles*, edited by V. A. Hackley and J. Texter (American Ceramic Society, Westerville, 1998), pp. 165–175.
[19] W. Han, "Viscothermal Coupling Effects on Sound Attenuation in Concentrated Colloidal Dispersions," Ph.D. dissertation, University of Maine, Orono, ME, 1995.
[20] J. E. Blue and E. G. McLeroy, "Attenuation of sound in suspensions and gels," J. Acoust. Soc. Am. **44**, 1145–1148 (1968).
[21] S. D. Richards, T. G. Leigthon, and N. R. Brown, "Visco-inertial absorption in dilute suspensions of irregular particles," Proc. R. Soc. London, Ser. A **459**, 2153–2167 (2003).
[22] W. S. Ament, "Sound propagation in gross mixtures," J. Acoust. Soc. Am. **25**, 683–641 (1953).
[23] R. L. Gibson and M. N. Toksöz, "Viscous attenuation of acoustic waves in suspensions," J. Acoust. Soc. Am. **85**, 1925–1934 (1989).
[24] A. H. Harker and J. A. G. Temple, "Velocity and Attenuation of Ultrasound in Suspension of Particles in Fluids," J. Phys. D **21**, 1576–1588 (1988).
[25] F. Babick, "Characterization of Submicron Emulsions by Ultrasonic Spectroscopy," Ph.D. dissertation, Technische Universität Dresden, Dresden, Germany, 2005 (in German), urn: http://nbn-resolving.de/urn:nbn:de:swb:14-1112809488055-07176
[26] Z. W. Qian, "Fractal dimensions of sediments in nature," Phys. Rev. E **53**(3), 2304–2306 (1996).

The interaction of ultrasound with particulate composites

Christopher Layman and N. Sanjeeva Murthy
Department of Physics, University of Vermont, Burlington, Vermont 05405

Ruey-Bin Yang
Department of Aerospace Engineering, Feng Chia University, Taichung, Taiwan

Junru Wu[a)]
Department of Physics, University of Vermont, Burlington, Vermont 05405

(Received 5 August 2005; revised 5 December 2005; accepted 6 December 2005)

The longitudinal and shear wave phase velocities and attenuation are measured as function of frequency for random particulate composites, consisting of near mono-disperse spherical glass particles imbedded in a homogeneous epoxy matrix. Experimental results are compared to two theoretical scattering models: the Waterman and Truell method (WT) and the dynamic generalized self-consistent model (DGSCM). It is observed that both models work well in the case of a low inclusion concentration. At higher concentrations both the WT and DGSCM matched the phase velocity, in most cases, within the range of experimental error of about 3%. The WT agrees with the attenuation data only at low concentrations, and the disparity between theory and measured values increase with increasing concentration. The DGSCM agrees well with the attenuation data. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2161450]

PACS number(s): 43.35.Cg, 43.35.Mr [YHB] Pages: 1449–1456

I. INTRODUCTION

Some polymeric composite materials consist of a viscoelastic matrix and randomly distributed imbedded inclusions of various sizes and shapes to improve their stiffness and strength characteristics. Ultrasound as a nondestructive technique to evaluate and characterize dynamic mechanical properties of such matrix/inclusions composite materials is quite useful. To achieve this goal, theories have been developed to address the interaction between ultrasonic waves and composite materials. A number of theories currently existing in the available literature predict the ultrasonic behavior of random particulate composites (see Anson and Chivers, 1993 for an overview of twelve different theories). They can generally be categorized into two groups: one that considers interactions among scattered waves (multiple scattering) generated by inclusions and one that does not. The treatment of the former usually involves a configurational averaging procedure among other things (Waterman and Truell, 1961; Yang, 2003; Kim, 2004) and the latter is based on the interaction of ultrasound with a single inclusion (IUSI); the correlations and interactions among inclusions are neglected (Brauner and Beltzer, 1988; Biwa, 2001).

The multiple scattering formalism developed by Waterman and Truell (WT) is based (Waterman and Truell, 1961; Foldy, 1945; Lax, 1952) on a configurational averaging of the scatterers' positions in the matrix. The WT method, in principle, has validity only for lower concentration composites because in this procedure the limit is taken of vanishing correlation between the positions of the scatterers (Waterman and Truell, 1961).

For composites with a higher concentration of inclusions an extension of the WT method was introduced to overcome the low concentration limit of WT. This procedure, termed the dynamic generalized self-consistent method (DGSCM), assumes an inclusion is surrounded by a coating of the matrix material; this coated sphere is then imbedded in an effective medium that has the overall properties of the real composite (Yang, 2003; Kim, 2004). A similar procedure is used in optics and is known as the Maxwell-Garnett effective medium theory (Niklasson *et al.*, 1981; Videen and Sun, 2003). To the authors' knowledge, the experimental verification of DGSCM is rather scarce (Kim *et al.*, 1995). In particular, the experimental study of the shear wave phase velocities and attenuations for composite materials is lacking.

The purpose of this work is to perform a more extensive study on the ultrasonic properties of viscoelastic particulate composites including phase velocity and attenuation for longitudinal as well as shear waves. Our approach is first to develop a well-defined composite material model of various inclusion concentrations by introducing spherical glass inclusions into an epoxy matrix. Second, ultrasonic spectroscopy (Wu, 1996) is used to measure phase velocity and attenuation as functions of frequency of longitudinal and shear waves for different inclusion concentrations. The experimental data obtained are then compared with WT and DGSCM.

II. THEORY

A. WT approach

The derivation of Waterman and Truell's configurational averaging procedure is beyond the scope of this paper and only the central results are reviewed. In the limit of zero correlation between the positions of the scatterers the effec-

[a)]Author to whom correspondence should be addressed; electronic mail: junru.wu@uvm.edu

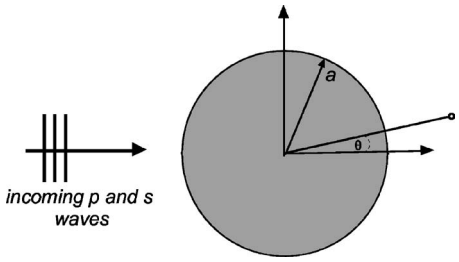


FIG. 1. Plane wave scattering off single sphere.

tive wave numbers for the composite material after applying an averaging technique can be expressed as (Waterman and Truell, 1961)

$$\left(\frac{\langle k\rangle}{k_m}\right)^2 = \left(1 + \frac{2\pi n_0 f[0]}{k_m^2}\right)^2 - \left(\frac{2\pi n_0 f[\pi]}{k_m^2}\right)^2, \tag{1}$$

$$\left(\frac{\langle \kappa\rangle}{\kappa_m}\right)^2 = \left(1 + \frac{2mn_0 g[0]}{\kappa_m^2}\right)^2 - \left(\frac{2\pi n_0 g[\pi]}{\kappa_m^2}\right)^2, \tag{2}$$

where $\langle k\rangle$ and $\langle \kappa\rangle$ are the longitudinal and shear effective wave numbers for the composite material, respectively, and the subscripts i and m on a property refer to the inclusion and matrix, respectively. Here $f[0]$ and $f[\pi]$ are the scattering amplitude functions in the forward and backward directions for an incident longitudinal wave. Likewise, for an incident shear wave the scattering amplitude functions are $g[0]$ and $g[\pi]$. The quantity n_0 is the number density of the scatterers. These scattering amplitude functions are obtained from the solution to the problem of the scattering of a plane wave off a single spherical particle (Fig. 1) in an infinite homogeneous medium (Ying and Truell, 1956; Einspruch *et al.*, 1960; Van de Hulst, 1957). The scattering amplitude functions are expressed as

$$f[\vartheta] = \frac{1}{ik_m}\sum_{n=0}^{\infty}(2n+1)A_n P_n[\cos\vartheta] \tag{3}$$

for an incident longitudinal wave and

$$g[\vartheta] = \frac{1}{i2\kappa_m}\sum_{n=0}^{\infty}(2n+1)(A_n + B_n)P_n[\cos\vartheta] \tag{4}$$

for an incident shear wave. Here A_n and B_n are the coefficients associated with the scattered waves, which are found by assuming an incident monochromatic plane wave and applying the appropriate boundary conditions and i is the unit for the imaginary number.

B. DGSCM approach

The physical model, on which the DGSCM is based, is shown in Fig. 2. Here a sphere of the inclusion material is surrounded by a coating of the matrix material. The volume concentration of the inclusions is chosen to correspond to the volume ratio of the two concentric spheres of radius a and b in the DGSCM model (Niklasson *et al.*, 1981):

$$c = \frac{a^3}{b^3}.$$

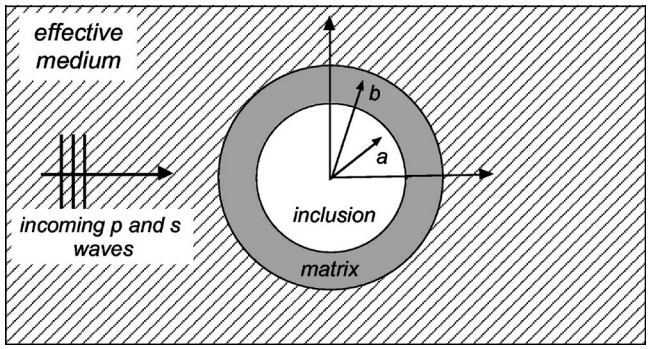


FIG. 2. Theoretical picture for the generalized self-consistent multiple scattering model (DGSCM).

This coated sphere is then imbedded in a hypothetical host medium, which has the properties of the real composite material. An incident wave consisting of plane longitudinal and transverse waves is then allowed to travel in this hypothetical medium and be incident upon the coated sphere. Equations (1) and (2) are still used, but now the k_m and κ_m should be replaced by $\langle k\rangle$ and $\langle \kappa\rangle$ (Yang, 2003), therefore:

$$1 = \left(1 + \frac{2\pi n_0 f[0]}{\langle k\rangle^2}\right)^2 - \left(\frac{2\pi n_0 f[\pi]}{\langle k\rangle^2}\right)^2, \tag{5}$$

$$1 = \left(1 + \frac{2\pi n_0 g[0]}{\langle \kappa\rangle^2}\right)^2 - \left(\frac{2\pi n_0 g[\pi]}{\langle \kappa\rangle^2}\right)^2. \tag{6}$$

These equations can be solved for the effective wave numbers by iteration methods and by assuming that the effective density of the composite can be written as

$$\langle \rho\rangle = \rho_i c + (1-c)\rho_m, \tag{7}$$

where ρ_m and ρ_i are the densities of the matrix and inclusion. The detailed mathematical description can be found in Yang's work (2003).

III. MATERIALS

A. Sample preparation

To manufacture samples of an arbitrary concentration (below that of a fully packed system) a simple technique is developed to significantly reduce the mobility of the glass particles within the matrix as it cures. Although other techniques also exist to accomplish this (Kinra *et al.*, 1980), the advantage of our technique is that it only requires a single pouring. The process begins by adding together in a small beaker an amount of epoxy and glass particles, which give the desired particle volume concentration. The sample volume V_{sam} is kept constant, so for a particular particle volume concentration c the mass of matrix material m_m and glass particles m_i are governed by: $m_m = V_{\text{sam}}(1-c)\rho_m$, and $m_i = V_{\text{sam}}(c)\rho_i$. It is also assumed that the particle arrangement within the matrix is uniform and approximately can be described like that of a simple cubic lattice. This allows for an expression relating the interparticle distance d and the particle concentration:

$$d = a\left(\frac{4\pi}{3c}\right)^{1/3}. \tag{8}$$

This combination is thoroughly mixed and put under a vacuum generated by a vacuum pump (Model SKH3SKG113F, GE, Princeton, NJ) for 15 min to remove all trapped air. Subsequent to the vacuuming process, the sample is put on a rolling stage, which keeps the glass particles evenly distributed. The sample is kept on the rolling stage for 20 min to allow the epoxy to just slightly begin to set, however it remains thin enough to pour. Meanwhile, an aluminum circular mold of radius 2.54 cm and thickness 0.23 cm is placed in a freezer, at a temperature of −15 °C, for a time long enough to reach the thermal equilibrium. The sample is taken off the rolling stage and is quickly poured into the cooled mold in the freezer. Once poured the epoxy's viscosity quickly rises to a point where the mobility of the glass particles is significantly reduced, thus preserving, as best as possible, the homogeneous distribution of the particles. The sample is left in the freezer for 24 h and then in a refrigerator, at a temperature of 5 °C, for another 24 h. After refrigeration, the sample is placed at room temperature (20 °C) for 24 more hours. After refrigeration, the epoxy is cured enough that if placed in a room temperature environment the glass particles do not move significantly. To ensure a total cure the sample is then placed in a 55 °C oven for 5 h. Thus, after this procedure a fully cured sample of an epoxy matrix with a random distribution of glass micro spheres, of any desirable concentration, is produced. Visually the spheres appear to be randomly and more and less evenly distributed, although small unevenness along vertical direction is unavoidable due to gravity. All the samples are then planned smooth to an even thickness in preparation for ultrasonic testing.

B. Constituent properties

The samples are prepared with glass micro-spheres (Duke Scientific Inc., Palo Alto, CA) of an average radius of 46 μm dispersed in a viscoelastic epoxy matrix (Tra-Cast 3012, Tra-Con Inc., Bedford, MA). The physical properties of the constituents are written in terms of their complex wave numbers:

$$k = \frac{\omega}{V_P[\omega]} + i\alpha_P[\omega], \tag{9}$$

$$\kappa = \frac{\omega}{V_s[\omega]} + i\alpha_s[\omega]. \tag{10}$$

Here V, α are the phase velocity and attenuation and λ and μ are the Lamé constants, respectively. The subscripts p and s refer to properties of longitudinal and shear waves, respectively. The properties of the glass micro-spheres (lime soda and borosilicate glass), except for the density $\rho_i = 2.5$ g/cm^3, are not available from the manufacturer. The values for the Lamé constants for the glass micro-spheres used are those listed in the work by Kinra *et al.* (1980), $\lambda_i = 25.8$ GPa and $\mu_i = 26$ GPa. The glass is assumed to be perfectly elastic (no attenuation) and to have a constant phase velocity in the frequency range used for these experiments. The wave numbers for the microspheres and the matrix, in the linear and isotropic theory of elasticity, are

$$k_i = \omega\sqrt{\frac{\rho_i}{\lambda_i + 2\mu_i}}, \tag{11}$$

$$\kappa_i = \omega\sqrt{\frac{\rho_i}{\mu_i}}, \tag{12}$$

$$k_m = \omega\sqrt{\frac{\rho_m}{\lambda_m + 2\mu_m}}, \tag{13}$$

$$\kappa_m = \omega\sqrt{\frac{\rho_m}{\mu_m}}. \tag{14}$$

The properties of the viscoelastic epoxy matrix are measured from the production of blank (no inclusions) epoxy samples. Four blank samples are made and each one is tested to obtain the phase velocity and attenuation profile as a function of frequency. These values are averaged and a curve (fourth-order polynomial for the velocity and linear fit for the attenuation) fitting is performed based on the average values. The functions found through curve-fitting for phase velocities and attenuations are used in Eqs. (7) and (8). These expressions are

$$V_p(f) = 2393.7795 + 40.6812f - 6.1494f^2 + 0.4617f^3 - 0.0132f^4, \tag{15}$$

$$V_s(f) = 1038.9408 + 31.9571f - 3.1019f^2 + 0.1076f^3 - 0.0005f^4, \tag{16}$$

$$\alpha_p(f) = -0.71 + 5.12f, \tag{17}$$

$$\alpha_s(f) = 14.20 + 24.13f, \tag{18}$$

for the longitudinal and shear phase velocity and attenuation, respectively. Evidently, there is weak dispersion in phase velocities both for the longitudinal and shear waves for the epoxy. In these equations f is the frequency in megahertz. Also, the velocity is in m/s and the attenuation is in dB/cm. The matrix's average density is $\rho_m = 1.176$ g/cm^3. Using these values Eqs. (1), (2), (5), and (6) are solved for the effective wave numbers of the composite and compared with the experimental results.

IV. EXPERIMENTS

A. Method

The shear and longitudinal wave data are obtained by the ultrasonic broadband substitution method. This technique, described in detail elsewhere (Wu, 1996), provides a means of determining both attenuation and phase velocity simultaneously over a range of frequencies. In brief, a broadband pulse generated by a 2.54-cm-diam PVDF transducer (Ultrasonic Science Ltd. Fleet, Hampshire, England) that is electronically excited by a pulser/receiver (Model PR5000, Matec instrument Companies, Inc., Northborough, MA) in a

water tank is received by the another similar PVDF transducer aligned along the wave propagation direction and digitized by a digital oscilloscope of 500-MHz bandwidth and 1-GHz sampling rate (Infinium, Hewlett Packard, CA). Then, a sample is inserted between them allowing the sound wave to propagate through the sample normally. The wave form transmitted through the specimen is digitized. A complex spectrum is obtained from the test sample wave form and divided by the complex spectrum of the reference water wave form without the sample presence to calculate the attenuation and velocity of the sample. Angular rotation of the sample with respect to the vertical axis in the sample plane that is perpendicular to the wave propagation direction is used to cause complete mode conversion to obtain the shear wave characteristics (Wu, 1996). Data are corrected for impedance discontinuities at the sample interfaces by Fresnel equations.

A single pair of PVDF broadband transducers are used in all the experiments. Their center frequency is determined to be about 6 MHz and the −6-dB bandwidth is about 10 MHz in water, giving a frequency range from about 1.5 to 13 MHz. For the longitudinal wave measurements the frequency range is only slightly reduced in the upper frequency components due to the presence of a sample, resulting in a range from about 1.5 to 10 MHz. The shear wave measurements produce a severe drop in the allowable frequencies due to the sample, the range is from 1.5 to 5 MHz.

B. Longitudinal data

Figures 3(a)–3(c) show the longitudinal phase velocity and attenuation as a function of frequency/*ka* for three different concentrations: 5.6%, 16%, and 41.3%. For the 5.6% concentration sample both theories match well with the data. As the particle concentration increases the differences in both phase velocity and attenuation predicted by WT and DGSCM increase. The WT method predicts the attenuation well at the low concentration case (at 5.6%, the values predicted by WT almost completely overlap with the experimental data of the composite), however it begins to overpredict as the concentration increases. The reasoning behind the overprediction by the WT method is that since the higher order statistics have been ignored in its derivation, the stronger interaction between adjacent particles that comes with increasing concentration has been omitted. This interaction manifests itself as a multiple scattering effect that tends to focus the effective scattered wave in the forward direction and hence lower the attenuation. In the figures, the attenuation due to the matrix along was also included for comparison. The difference between attenuation of the composite and that due to the matrix alone should approximately indicate the contribution of scattering by the inclusions to the attenuation. (It should be cautious that since the matrix attenuation is obtained from the blank epoxy samples, no inclusions are included. The data have not been scaled down due to the presence of the inclusions, so the difference does not equal to the contribution of scattering.)

Figures 4(a) and 4(b) show the longitudinal velocity and attenuation as a function of particle concentration for two

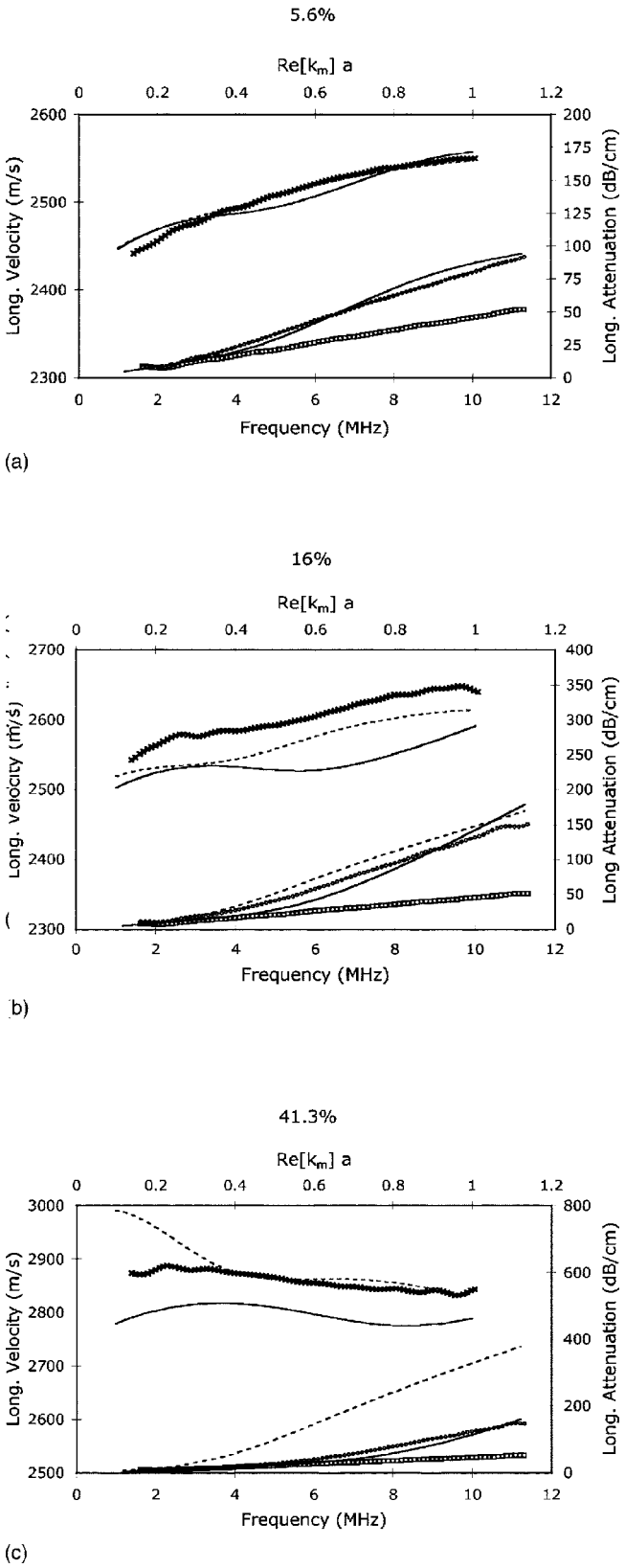
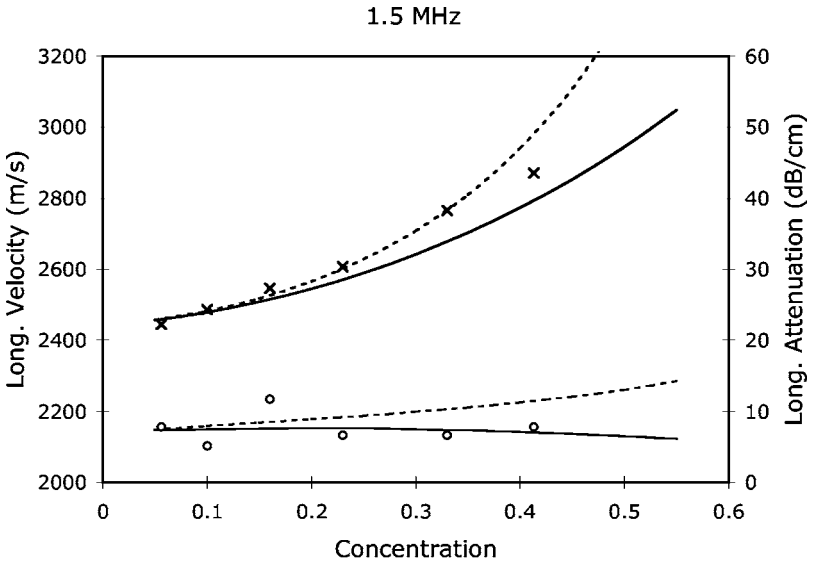
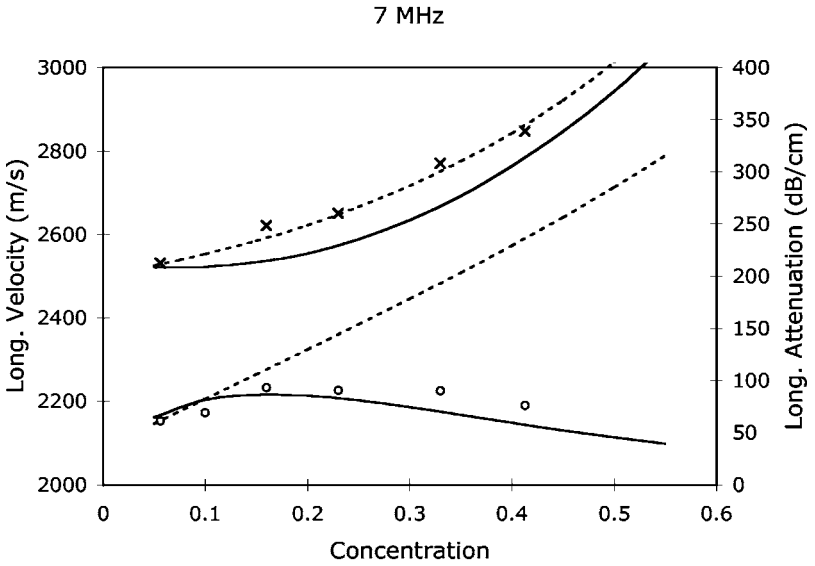


FIG. 3. (a) Longitudinal phase velocity and attenuation vs frequency (bottom horizontal axis)/*ka* (top horizontal axis) for particle concentration of 5.6%. Note that the experimental attenuation data of the composite almost totally overlap with predicted data by WT. (b) Longitudinal phase velocity and attenuation vs frequency (bottom horizontal axis)/*ka* (top horizontal axis) for particle concentration of 16%. (c) Longitudinal phase velocity and attenuation vs frequency (bottom horizontal axis)/*ka* (top horizontal axis) for particle concentration of 41.3%. (×) Experimental velocity data. (○) Experimental attenuation data. Solid line: DGSCM. Dashed line: WT. (□) Experimental attenuation due to matrix alone.



(a)



(b)

FIG. 4. (a) Longitudinal phase velocity and attenuation vs concentration at 1.5 MHz. (b) Longitudinal phase velocity and attenuation vs concentration at 7 MHz. (×) Experimental velocity data. (○) Experimental attenuation data. Solid line: DGSCM. Dashed line: WT.

different frequencies: 1.5 and 7 MHz. Both theories match the velocity data well for both frequencies, albeit the WT model seems slightly better in the higher frequency than in the lower one (for higher concentrations). For the attenuation at 1.5 MHz both theories match the data well at all concentrations. However, at the higher frequency the WT model ceases to be valid at a concentration of about 15%, while the DGSCM is significantly closer to the experimental data.

C. Shear data

As stated earlier, the higher attenuation associated with propagation of shear waves reduces the frequency range of the data significantly. Figures 5(a)–5(c) show the shear wave velocity and attenuation as a function of frequency. For the 5.6% and 16% samples both methods reasonable match the velocity data. For 41.3% the WT method is in disagreement, while agreement of the DGSCM is fairly well. The shear attenuation follows the same trend as in the longitudinal case, the WT matches well in the lower concentration case and overpredicts in the higher concentrations. The DGSCM at these higher concentrations seems to agree well. Once again, in the figures, the attenuation due to the matrix alone is also included for comparison.

Figures 6(a) and 6(b) show the shear velocity and attenuation as a function of particle concentration for the two frequencies 1.5 and 5 MHz. The same trend as in the longitudinal velocity case is seen here, where the WT seems to match better at the higher frequency and worse in the lower (for higher concentrations). Likewise, for the shear attenuation the same trend is seen as in the longitudinal case. Both theories work well in the lower frequency case, while at the higher frequency the WT model diverges from the experimental data at approximately 10% concentration.

V. CONCLUSIONS AND DISCUSSIONS

The longitudinal and shear wave phase velocities and attenuation are measured for random particulate composites, consisting of spherical glass particles imbedded in a homo-

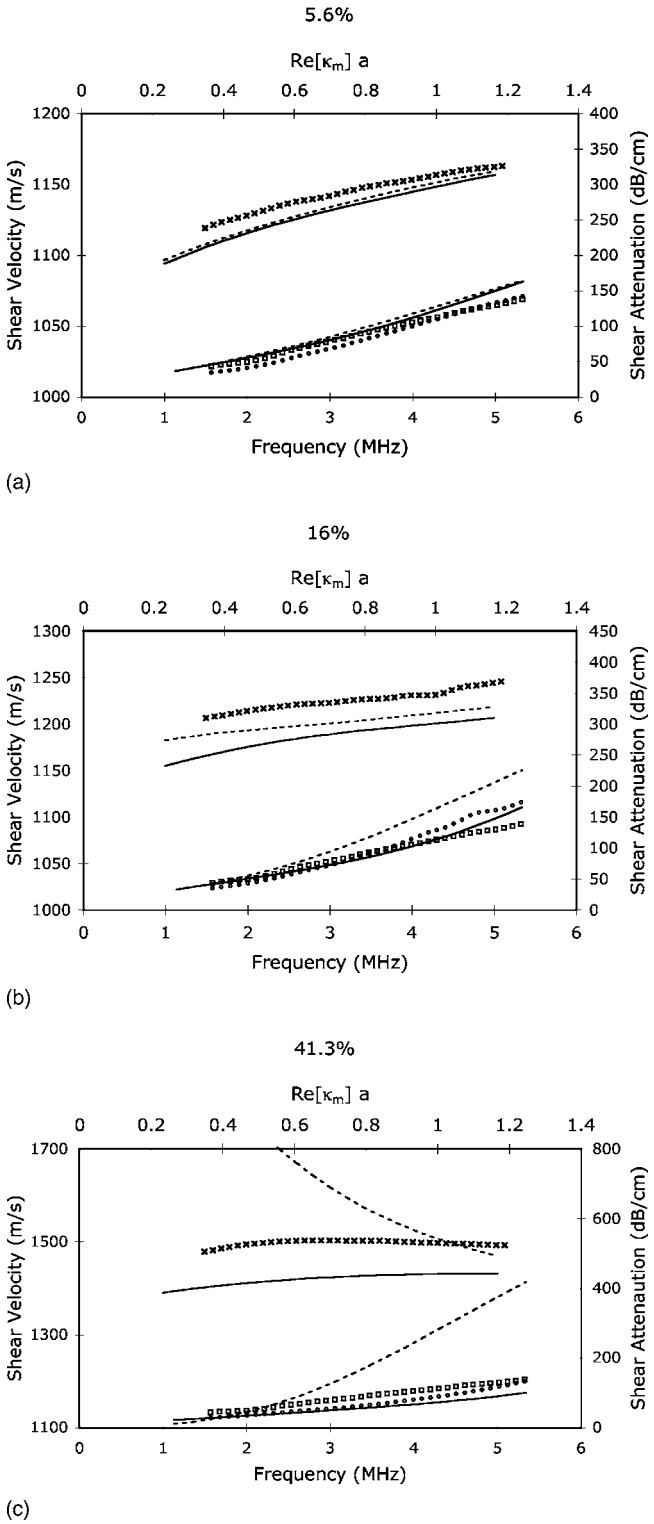


FIG. 5. (a) Shear phase velocity and attenuation vs frequency (bottom horizontal axis)/*ka* (top horizontal axis) for particle concentration of 5.6%. (b) Shear phase velocity and attenuation vs frequency (bottom horizontal axis)/*ka* (top horizontal axis) for particle concentration of 16%. (c) Shear phase velocity and attenuation vs frequency (bottom horizontal axis)/*ka* (top horizontal axis) for particle concentration of 41.3%. (×) Experimental velocity data. (○) Experimental attenuation data. Solid line: DGSCM. Dashed line: WT. (□) Experimental attenuation due to matrix alone.

geneous epoxy matrix. These measurements are then compared to two theoretical scattering models: the Waterman and Truell method and the dynamic generalized self-consistent model. It is seen that both models worked well in the case of a low inclusion concentration. At higher concentrations both the WT and DGSCM match the phase velocity, in most cases, within the range of experimental error (see the following for error analysis) of about 3%. On the other hand, the WT agrees with the attenuation data only at low concentrations, and its disagreement increases with increasing concentration. In general, the DGSCM agrees better with the attenuation data than the WT method; the same conclusion was reached by the study performed by Kim *et al.* (1995) for longitudinal waves.

It is of interest to note the changes in the overall physical properties that occur in these samples as the glass particles are introduced. It is first assumed that the physical properties of the composites can be summarized in the form of the Lamé constants λ and μ and mass density ρ, respectively. These constants are related to the longitudinal and shear wave speeds by Eqs. (11) and (12) (in the linear and isotropic theory of elasticity). Both WT and DGSCM assume that ρ of the composites is determined by Eq. (7) and does not change with frequency. In all the phase velocity versus concentration figures it is observed that phase velocity increases with increasing concentration. The degree of increase of phase velocity with concentration seems to be sensitive to frequency. The Lamé constants, and hence the dynamic physical properties of the samples, will follow this same trend. As an example we compare the Lamé constants for the concentrations of 5.6% and 41.3%, at frequency of 1.5 MHz. At concentration 5.6% the λ is 4.3 GPa and μ is 1.6 GPa, and at 41.3% concentration λ is 6.7 GPa and μ is 3.8 GPa. As can be seen here, the increase in both λ and μ from 5.6% concentration to 41.3% concentration is dramatic, there is a 57% and 237% increase, respectively. On the contrary, in the frequency range tested, attenuation is much less sensitive to concentrations.

For longitudinal waves, it is observed from Figs. 3(a)–3(c) that the attenuation due to scattering of inclusions increases monotonically as frequency increases. For the shear waves, it is quite different. From Figs. 5(a)–5(c), it is seen that in the some frequency range, the attenuation of composites actually is smaller than that of matrix alone. It is mainly due to very high attenuation of matrix shear waves. The presence of inclusions substitutes a portion of sample volume occupied by the matrix of high attenuation by that of inclusions of negligible attenuation. Additionally, high attenuation of matrix would reduce the amplitude of shear wave in the sample dramatically and thus decrease the effect of scattering.

Concerning the error associated with the collection of the experimental data, there are at least three sources: time measurements taken from the oscilloscope, unevenness of the thickness of the samples, and the density of the samples. A modern digital oscilloscope like the one we used measure time very accurately; its associated error should be negligible in our case. The unevenness of the thickness of the samples was a result of the milling procedure during a sample preparation. To calculate this error several measurements are made at various places on a single sample. It is determined from these measurements that the error in the thickness is less than

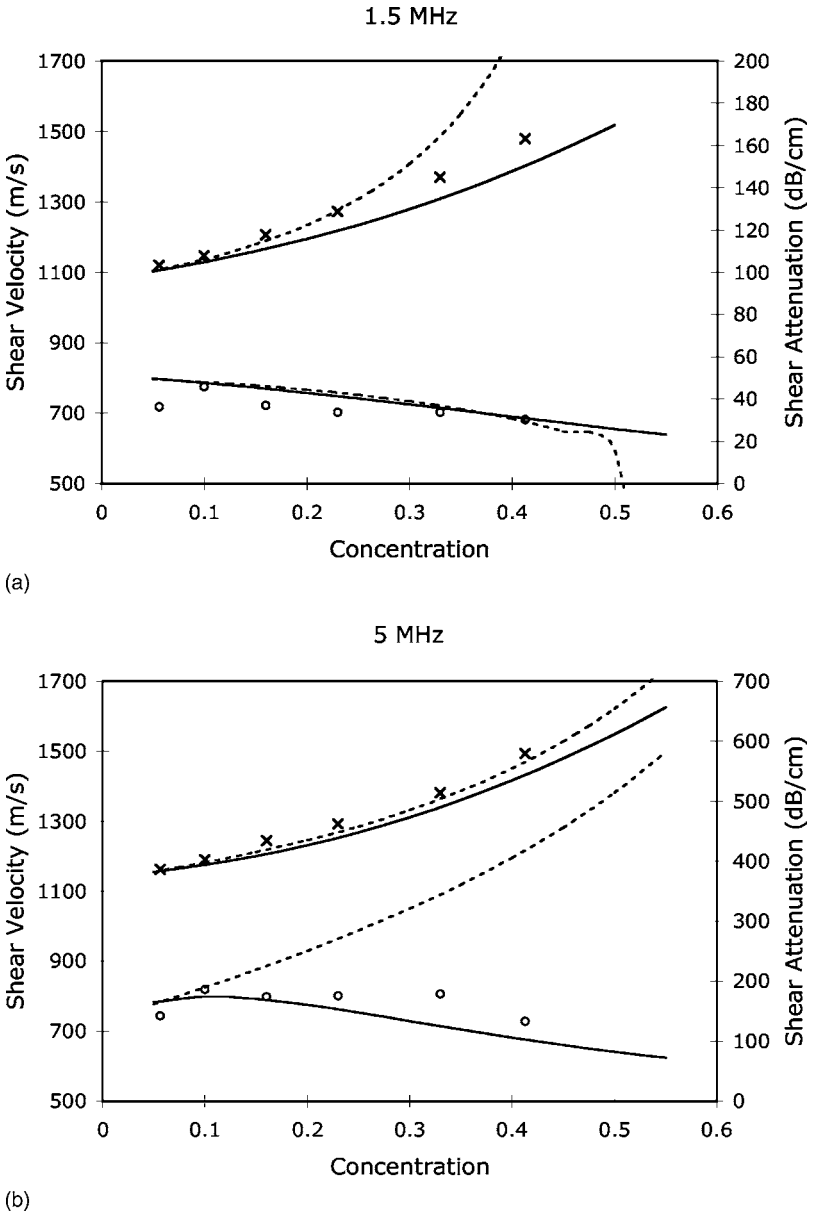


FIG. 6. (a) Shear phase velocity and attenuation vs concentration at 1.5 MHz. (b) Shear phase velocity and attenuation vs concentration at 5 MHz. (×) Experimental velocity data. (○) Experimental attenuation data. Solid line: DGSCM. Dashed line: WT.

1%. The density measurements are made using Archimedes's method; this is a very accurate method and hence any error associated with it is negligible.

Another source of error is associated with diffraction loss. In our previous publication (Wu, 1996), it is calculated that the error is less than 0.5%.

On the theoretical side there is also some uncertainty as to what values to use for the concentration and particle size. The size distribution of the particles is determined by the measurement of 230 particles. The mean size is 46 μm (radius) and the standard deviation is 2.8 μm. The size distribution of this group of particles fit a normal distribution curve. Several calculations (at different concentrations) of the velocity and attenuation are made using first the mean radius and then using the mean radius plus or minus one standard deviation. It is seen over a wide range of frequencies that the change in the velocity, due to this shift in particle size, is negligible. On the other hand, more significant changes in the attenuation are seen. In addition, these changes are seen not to be uniform over the frequencies used, i.e., a 5% change in the radius of the particle resulted in a variable percent change in the attenuation over the frequency range used. As stated earlier, it is observed that the distribution of the particles along the vertical direction within the matrix is not as uniform as we would like; this might also introduce errors. For the overall uncertainty involved in the comparison between the experimental data and the theoretical simulations we give the estimate of about 3%.

ACKNOWLEDGMENT

C.L. was supported by a research assistantship from NSF/VT EPSCoR.

Anson, L. W., and Chivers, R. C. (**1993**). "Ultrasonic velocity in suspensions of solids in solids-A comparison of theory and experiment," J. Phys. D **26**, 1566–1575.

Biwa, S. (**2001**). "Independent scattering and wave attenuation in viscoelastic composite," Mech. Mater. **33**, 635–647.

Brauner, N., and Beltzer, A. I. (**1988**). "Wave-obstacle interaction in a lossy medium: Energy perturbations and negative extinction," Ultrasonics **26**, 328–334.

Einspruch, N. G., Witterholt, E. J., and Truell, R. (**1960**). "Scattering of a plane transverse wave by a spherical obstacle in an elastic medium," J. Appl. Phys. **31**, 806–818.
Foldy, L. L. (**1945**). "The multiple scattering of wave. I. General theory of isotropic scattering randomly distributed scatterers," Phys. Rev. **67**, 107–119.
Kim, J. Y. (**2004**). "On the generalized self-consistent model for the elastic wave propagation in composite materials," Int. J. Solids Struct. **41**, 4349–4360.
Kim, J. Y., Ih, J. G., and Lee, B. H. (**1995**). "Dispersion of elastic waves in random particulate composites," J. Acoust. Soc. Am. **97**, 1380–1388.
Kinra, K. V., Petraitis, S. M., and Datta, K. S. (**1980**). "Ultrasonic wave propagation in a random particulate composite," Int. J. Solids Struct. **16**, 301–312.
Lax, M. (**1952**). "Multiple scattering of waves. II. The effective field in dense systems," Phys. Rev. **85**, 621–629.
Niklasson, G. A., Granqvist, C. G., and Hunderi, O. (**1981**). "Effective medium models for the optical properties of inhomogeneous materials," Appl. Opt. **20**, 26–30.
Van De Hulst, H. C., *Light Scattering by Small Particles* (Wiley, New York, **1957**).
Videen, G., and Sun, W. (**2003**). "Yet another look at light scattering from particles in absorbing media," Appl. Opt. **42**, 6724–6727.
Waterman, P. C., and Truell, R. (**1961**). "Multiple scattering of waves," J. Math. Phys. **2**, 512–537.
Wu, J. (**1996**). "Determining of velocity and attenuation of shear waves using ultrasonic spectroscopy," J. Acoust. Soc. Am. **99**, 2871–2875.
Yang, R. B. (**2003**). "A dynamic generalized self-consistent model for wave propagation in particulate composite," J. Appl. Mech. **70**, 575–582.
Ying, C. F., and Truell, R. (**1956**). "Scattering of a plane longitudinal wave by a spherical obstacle in an isotropically elastic solid," J. Appl. Phys. **27**, 1086–1097.

Relationship between Nusselt number and the thermoviscous (Rott) functions

Jin Liu and Steven L. Garrett[a)]
Graduate Program in Acoustics, Penn State University, State College, Pennsylvania 16804

(Received 29 July 2005; revised 3 December 2005; accepted 13 December 2005)

The characterization of the thermal properties of regenerator materials has traditionally been studied by measurement of the heat transport of a gas passing through a regenerator flanked by two heat exchangers. The results for heat transfer between the gas and regenerator are typically expressed as a nondimensional Nusselt number. The newer thermoacoustic perspective, developed over the past two decades by Swift and his collaborators, has characterized the thermal properties of regenerators through the thermoviscous function used by Rott in the 1960s. This paper derives the relation between those two perspectives. The complex Nusselt number for oscillatory flow is calculated in terms of the thermoviscous f function. The asymptomatic results of the real part of the complex Nusselt number at zero frequency converge to the steady (direct current) result and at high frequency it matched qualitatively with the limit derived by other researchers interested in the behavior of heat exchangers used in oscillating flows. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2165000]

PACS number(s): 43.35Ud, 43.20Mv, 43.20Ks [RR] Pages: 1457–1462

I. INTRODUCTION

The use of a porous material as a medium for temporary thermal energy storage (i.e., a regenerator) has been popular since Rev. Stirling[1] first patented his "economizer" in the early 19th century. The development of the first thermoacoustic-Stirling engine that demonstrated competitive efficiency[2] has renewed interest in the characterization and optimization of regenerator materials, particularly at the frequencies of interest for thermoacoustic engines and refrigerators.

Regenerators in Stirling machines (engines or refrigerators) are intended to provide sufficient thermal contact to the working fluid (gas) to minimize loss due to irreversible heat transfer (nonisothermality losses) while generating as little viscous dissipation as possible.[3] This paper concentrates on regenerator thermal properties. (Our approach to characterization of the viscous dissipation has been described previously.[4])

There have been two approaches taken to the characterization of the thermal properties of regenerator materials. Since there is gas flow through the regenerator in a Stirling machine, it is natural to introduce a convective heat transfer coefficient, h, that relates the rate of heat transfer, $\dot{Q}$, per unit surface area, S, to the temperature difference between the solid matrix, T_{reg}, and the mean gas temperature, $\langle T_{\mathrm{gas}}\rangle$:

$$\frac{\dot{Q}}{S} = h(T_{\mathrm{reg}} - \langle T_{\mathrm{reg}}\rangle). \tag{1}$$

In this paper, the spatial-averaged temperature, $\langle T_{\mathrm{gas}}\rangle = 1/A\int T_{\mathrm{gas}}dA$, is used as the mean gas temperature instead of the bulk (velocity-weighted section averaged) temperature.

a)Electronic mail: sxg185@psu.edu

II. CONVECTIVE HEAT-TRANSFER COEFFICIENT

The local convective-heat-transfer coefficient may be related to the temperature gradient in the fluid at the surface of the solid. The exact expression of Fourier's Law of Conduction states that heat transfer is proportional to the temperature gradient at the wall

$$\frac{\dot{Q}}{S} = -k\left(\frac{\partial T_{\mathrm{gas}}}{\partial y}\right)_{y=0}. \tag{2}$$

Equating the right-hand side (rhs) of Eqs. (1) and (2),

$$h = -\frac{k}{(T_{\mathrm{reg}} - \langle T_{\mathrm{gas}}\rangle)}\left(\frac{\partial T_{\mathrm{gas}}}{\partial y}\right)_{y=0}. \tag{3}$$

The Nusselt number, Nu_r, is the dimensionless gas temperature gradient at the wall

$$\mathrm{Nu}_r = \frac{hr_h}{k} = \frac{r_h}{\langle T_{\mathrm{gas}}\rangle - T_{\mathrm{reg}}}\frac{\partial(T_{\mathrm{gas}} - T_{\mathrm{reg}})}{\partial y}. \tag{4}$$

The convective heat transport coefficient has been converted to a nondimensional Nusselt number that introduces a scale length which allows comparison to the thermal conductivity of the stationary gas, k. The scale length is usually chosen to reflect the ratio of the cross-sectional area of a single pore, A_{pore}, to the wetted perimeter of the pore, Π, that is designated as the hydraulic radius, $r_h = A_{\mathrm{pore}}/\Pi$. The most common choice for this length scale in the literature on heat exchange is the hydraulic diameter, $D_h = 4r_h$, therefore $\mathrm{Nu}_D = 4\mathrm{Nu}_r$. In this study we chose to use the hydraulic radius to be consistent with the thermoviscous f function which is defined as a function of the ratio of the hydraulic radius to the penetration depth, defined later.

Measurement of Nu in oscillatory flow is difficult. An apparatus that can be pressurized and provide measurable oscillatory gas flow must be fitted with an electrical heater to

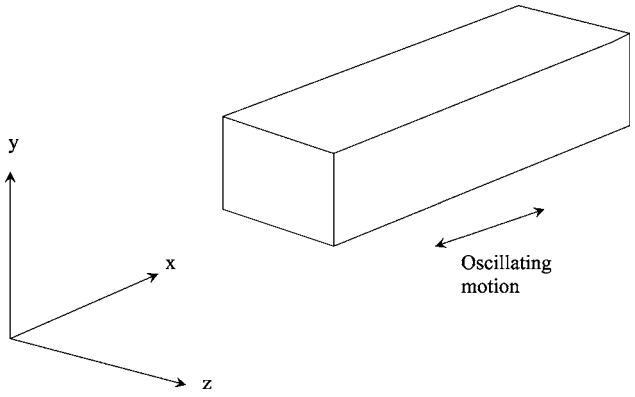


FIG. 1. The coordinate system.

apply a known heat load at one end of the regenerator and a heat exchanger at the other end of the regenerator is required to measure the heat leaving the regenerator.[5,6] Care must be taken to ensure that other heat flow paths between the heater and heat exchanger be eliminated or calibrated so that they can be separated from the heat flow of interest.

More recently, a new thermoacoustical perspective has been developed by Swift[7,8] that introduces a spatial-averaged thermoviscous complex function, f, to characterize the drag and heat transfer in porous media.

Following Swift's notation, using the coordinate systems shown in Fig. 1, and using Rott's acoustic approximation (the time dependent variables are small compared to their mean values, the penetrations depths are smaller than acoustic wavelengths, and only ideal gases are considered), the steady-state sinusoidal variables such as pressure, p, temperature, T, velocity, $\mathbf{v}$, and density, ρ, can be expanded in the complex notation, where $\omega=2\pi f$ is the angular frequency of oscillation. The subscript m stands for the mean value

$$p(x,y,z,t) = p_m(x) + \mathrm{Re}[p_1(x,y,z)e^{i\omega t}], \tag{5}$$

$$\mathbf{v}(x,y,z) = \mathrm{Re}[\mathbf{v}_1(x,y,z)e^{i\omega t}], \tag{6}$$

$$\rho(x,y,z,t) = \rho_m(x) + \mathrm{Re}[\rho_1(x,y,z)e^{i\omega t}], \tag{7}$$

$$T(x,y,z,t) = T_m(x) + \mathrm{Re}[T_1(x,y,z)e^{i\omega t}]. \tag{8}$$

For one-dimensional flow, the momentum equation becomes

$$i\omega\rho_m u_1 = -\frac{dp_1}{dx} + \mu\left[\frac{\partial^2 u_1}{\partial y^2} + \frac{\partial^2 u_1}{\partial z^2}\right], \tag{9}$$

where μ is the shear viscosity and $u(x,y,z,t) = \mathrm{Re}[u_1(x,y,z)e^{i\omega t}]$ is the x component of the velocity. Assuming the accessible solid heat capacity per unit volume is much larger than the fluid's heat capacity per unit volume inside the pore, the acoustic approximation for the general equation of heat transfer becomes

$$\rho_m c_p\left(i\omega T_1 + u_1\frac{dT_m}{dx}\right) - i\omega p_1 = k\left[\frac{\partial^2 T_1}{\partial y^2} + \frac{\partial^2 T_1}{\partial z^2}\right], \tag{10}$$

where c_p is the isobaric heat capacity per unit mass. After taking the spatial averages, $1/A\int[\ ]dA$, the averaged solution for $\langle u_1\rangle$ and $\langle T_1\rangle$ are

$$\langle u_1\rangle = \frac{1}{i\omega\rho_m}(1-f_\nu)\frac{dp_1}{dx}, \tag{11}$$

$$\langle T_1\rangle = \frac{1}{\rho_m c_p}(1-f_\kappa)p_1 - \frac{1}{i\omega A}\frac{dT_m}{dx}\frac{(1-f_\kappa)-\sigma(1-f_\nu)}{(1-f_\nu)(1-\sigma)}U_1, \tag{12}$$

where $U_1=\int u dA$ is the volumetric velocity and σ is the Prandtl number. Without a mean temperature gradient, dT_m/dx, $\langle T_1\rangle$ becomes,

$$\langle T_1\rangle = \frac{1}{\rho_m c_p}(1-f_\kappa)p_1. \tag{13}$$

f_κ and f_ν are the thermoviscous f functions (also known as the Rott functions and thermoacoustic f functions). They are complex functions which depend on the ratio of the hydraulic radius of pores to the relevant penetration depth: f_κ depends on r_h/δ_κ and f_ν depends on r_h/δ_ν. $\delta_\kappa=\sqrt{2k/\omega\rho c_p}$ is the thermal penetration depth and $\delta_\nu=\sqrt{2\mu/\rho\omega}$ is the viscous penetration depth. The thermoviscous f functions have been calculated for many geometries such as parallel plates, rectangular pores, and circular pores.[7] For stacked screens and other nonuniform geometries, the f functions are only approximate or must be determined empirically.

As shown in Eq. (13), the heat transfer between the fluid and the pores causes phase shifts between $\langle T_1\rangle$ and p_1 that are characterized by the thermal f_κ function. It is not easy to measure the averaged oscillating temperature, $\langle T_1\rangle$, inside the pore. Instead of measuring p_1 and $\langle T_1\rangle$ directly, the complex ratio of the displacement to pressure, ξ_1/p_1, can be measured to determine the f_κ function for different geometries using the technique described by Wilen.[9]

The first-order equation-of-state

$$\langle\rho_1\rangle = \frac{-\rho_m}{T_m}\langle T_1\rangle + \frac{\rho_m}{p_m}p_1, \tag{14}$$

can be combined with the continuity equation

$$i\omega\langle\rho_1\rangle + \frac{d}{dx}(\rho_m\langle u_1\rangle) = 0, \tag{15}$$

and Eq. (13) to produce an expression for the volume velocity, U_1, in terms of the f_κ function

$$dU_1 = -\frac{i\omega A dx}{\gamma p_m}[1+(\gamma-1)f_\kappa]p_1. \tag{16}$$

That produces the equivalent of a complex frequency-dependent gas compliance. Therefore, the ratio of the gas displacement to the oscillating pressure, ξ_1/p_1, equals

$$\frac{\xi_1}{p_1} = -\frac{L_{\mathrm{eff}}}{\gamma p_m}[1+(\gamma-1)f_\kappa]. \tag{17}$$

The effective length, L_{eff}, can be found by extrapolating the experimental results for ξ_1/p_1 to zero frequency, $\omega=0$. At zero frequency, the oscillation of the fluids inside the pore is isothermal so that

$$\frac{\xi_1(\omega=0)}{p_1(\omega=0)} = -\frac{L_{\mathrm{eff}}}{p_m}. \tag{18}$$

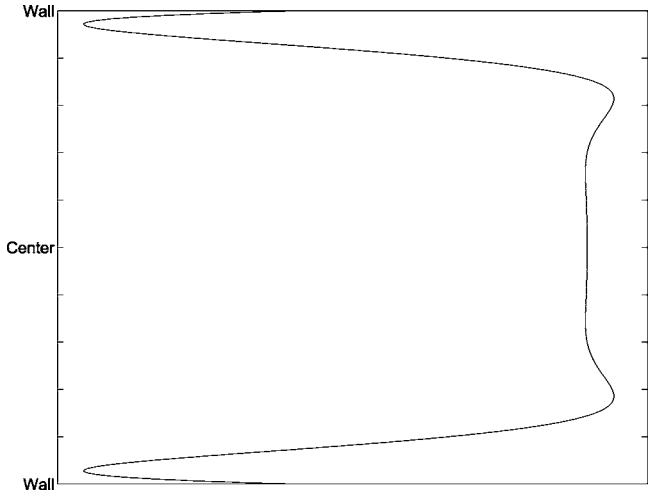


FIG. 2. Instantaneous temperature profile of T_1 for oscillating flow with oscillating pressure inside a circular pore. The hydraulic radius of the pore is five times the thermal penetration depth.

Using this method to investigate the thermal behavior of the oscillating flow, the experiment should be conducted at low velocity to minimize the viscous loss effect. Therefore, the f function measured by this method is limited to the laminar flow regime. In next section, the complex Nusselt number will be derived in terms of the f function. The solutions for the f function in simple geometries based on unidirectional laminar flow assumptions will be used.

III. COMPLEX NUSSELT NUMBER

In this section, the relationship between the Nusselt number, Nu, and the thermoviscous f function is derived. This provides an acoustic method to characterize the heat transfer behavior of porous media without requiring the measurement of the temperature and without requiring heat exchangers.

As shown in Eq. (4), the Nusselt number is the dimensionless temperature gradient. In steady [direct current (dc)] flow, the gas temperature gradient at the wall, $\partial T_{\mathrm{gas}}/\partial y|_{(y=0)}$ is proportional to and in-phase with the bulk gas-wall temperature difference, $\langle T_{\mathrm{gas}}\rangle - T_{\mathrm{reg}}$. Therefore, the heat transfer coefficient and Nusselt number are real numbers and Nusselt number is a real constant for fully developed internal flows.

But for oscillating flow, the bulk gas-wall temperature difference is not always in-phase with the wall temperature gradient. An instantaneous temperature profile of T_1 for oscillating flow inside circular pore is plotted in Fig. 2 using the solution given by Swift's Eq. (4.59).[7] In this plot, the hydraulic radius is five times the thermal penetration depth. Two important features appear in this plot: First, the center of the fluid far from the walls exhibits "plug flow;" the bulk temperature-wall gradient occurs mostly in a thin layer of fluid close to the wall. Therefore, the thinner the thermal penetration depth, the greater the temperature gradient. Second, all of the fluid inside the pore is not oscillating in-phase; the thin layer of the fluid next to the wall is leading the phase of the plug flow at the center. Because of these features of oscillating flow, it is not appropriate to use the heat transfer coefficient calculated or measured for steady-state flow. A complex heat transfer coefficient, and thus a complex Nusselt number, is required to account for the phase-shifts in the flow.

The literature on complex Nusselt numbers is sparse. Kornhauser and Smith[10] experimentally studied the complex Nusselt number for oscillating pressure and oscillating flow by measuring the surface heat flux and gas wall temperature differences. Later Kornhauser and Smith[11] did a similar study of complex Nusselt number from pressure-volume measurements only: instantaneous surface-averaged heat transfer and mixed mean gas temperature were calculated from the pressure, volume, and time data. Gedeon[12] calculated the exact solution for complex Nusselt number in incompressible flow between parallel plates based on a mean-parameter model.

A. Complex Nusselt number for heat transfer due to compression and expansion in oscillating flow

Consider the general equation of heat transfer for an inviscid ideal gas[13]

$$\rho c_p \frac{\partial T}{\partial t} - \frac{\partial p}{\partial t} + \rho c_p \mathbf{v}\cdot\boldsymbol{\nabla} T - \mathbf{v}\cdot\boldsymbol{\nabla} p = \boldsymbol{\nabla}\cdot(k\,\boldsymbol{\nabla} T). \tag{19}$$

This is the exact equation from the microscopic view in the fluid-filled space. The $\mathbf{v}\cdot\boldsymbol{\nabla} p \propto uKp$ term is small compared to $\partial p/\partial t \propto \omega p$ due to the fact that $\omega/K = c \gg v$, where c is the sound speed, u is the particle velocity, and K is the wave vector. In the absence of steady (dc) flow or a static temperature gradient, the linearized version Eq. (19) can be reduced to

$$\rho c_p \frac{\partial T}{\partial t} - \frac{\partial p}{\partial t} = \boldsymbol{\nabla}\cdot(k\,\boldsymbol{\nabla} T). \tag{20}$$

Next, we perform a local spatial average of Eq. (20) over the gas volume inside the pore of length dx and cross-sectional area A.[14] On the rhs, the divergence theorem is used to express the volume integral as the divergence of the heat flux density $k\boldsymbol{\nabla} T$ as a surface integral over the gas-solid wetted area S within the volume of integration. The result is

$$Adx\left(\left\langle \rho c_p \frac{\partial T}{\partial t}\right\rangle - \frac{\partial p}{\partial t}\right) = \int k\,\boldsymbol{\nabla} T dS. \tag{21}$$

Defining the heat transfer coefficient in the presence of an oscillating pressure to be h_p, and using the definition of hydraulic radius r_h, which is the ratio of the gas volume to gas-wetted surface area, we obtain

$$\left\langle \rho c_p \frac{\partial T}{\partial t}\right\rangle - \frac{\partial p}{\partial t} = -\frac{h_p}{r_h}(\langle T\rangle - T_m). \tag{22}$$

Using Rott's approximation, $T(x,y,z,t) = T_m(x) + \mathrm{Re}[T_1(x,y,z)e^{i\omega t}]$, and substituting Eq. (8) into Eq. (22),

$$\left\langle \rho c_p \frac{\partial T_1}{\partial t}\right\rangle - \frac{\partial p}{\partial t} = -\frac{h_p}{r_h}\langle T_1\rangle. \tag{23}$$

Solving for $\langle T_1\rangle$ and substituting into equation Eq. (13), the complex heat transfer coefficient can be written in term of thermal f_κ function

$$h_p = i\omega\rho_m c_p r_h \left(\frac{f_\kappa}{1 - f_\kappa} \right). \tag{24}$$

Using the definition of the square of the thermal penetration depth,

$$\delta_\kappa^2 = \frac{2k}{\rho c_p \omega}, \tag{25}$$

h_p can be expressed in terms of the penetration depth as

$$h_p = \frac{2ikr_h}{\delta_\kappa^2} \left(\frac{f_\kappa}{1 - f_\kappa} \right). \tag{26}$$

When the penetration depth is much smaller than the hydraulic radius of the pores, the boundary layer limits for the thermoviscous f function are[7]

$$f_\kappa \approx (1 - i)\frac{\delta_\kappa}{2r_h}, \tag{27}$$

$$f_\nu \approx (1 - i)\frac{\delta_\nu}{2r_h}. \tag{28}$$

Substituting Eq. (27) into Eq. (26) and taking the high-frequency limit

$$\lim_{\omega \to \infty} \frac{\delta_\kappa}{r_h} = 0, \tag{29}$$

the boundary layer limit of h_p becomes

$$\lim_{\omega \to \infty} h_p = \frac{k}{\delta_\kappa}(1 + i). \tag{30}$$

The complex Nusselt number for oscillating flow, subject to oscillating pressure, is Nu_p,

$$\mathrm{Nu}_p = \frac{h_p r_h}{k} = 2i(Lc_\kappa)^2 \left(\frac{f_\kappa}{1 - f_\kappa} \right), \tag{31}$$

where the Lautrec number is $\mathrm{Lc}_\kappa \equiv r_h/\delta_\kappa$. The boundary layer limit for Nu_p is

$$\lim_{\omega \to \infty} \mathrm{Nu}_p = \frac{r_h}{\delta_\kappa}(1 + i) = (1 + i)\mathrm{Lc}_\kappa. \tag{32}$$

The complex Nusselt numbers for circular pores, square pores, and parallel plates, subject to oscillating pressure, are plotted in Fig. 3. The boundary layer limit, based on Eq. (32), is plotted in black diamond symbols. The real part of the Nusselt number is much larger than the imaginary part at low Lautrec numbers. That means the special averaged temperature is in-phase with the gas-wall temperature gradient when the penetration depth is much larger than the pore size, as it would be in a regenerator.

B. Complex Nusselt number for oscillating flow subject to a temperature gradient

The dimensionless form of Eq. (19) is

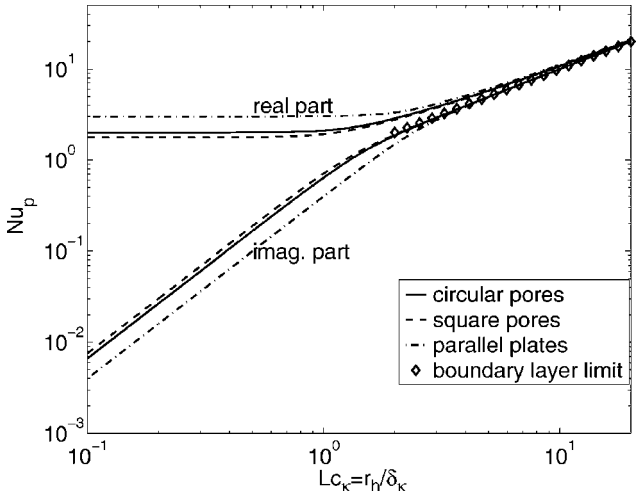


FIG. 3. Complex Nusselt number for circular pores, square pores, and parallel plates subject to oscillating pressure. The black diamond symbols are the boundary layer limit from Eq. (32).

$$\frac{L}{u/i\omega} T^* + u^* \frac{\partial T^*}{\partial x^*} - \frac{i\omega u_m L}{c_p T_m} p^* - \frac{u_m^2}{c_p T_m} u^* \frac{\partial p^*}{\partial x^*} = \frac{k/r_h}{\rho c_p u_m} \frac{L}{r_h} \frac{\partial^2 T^*}{\partial y^{*2}},$$
$$u^* = \frac{u}{u_0}, \quad T^* = \frac{T}{T_m}, \quad p^* = \frac{p}{\rho u_0^2}. \tag{33}$$

L is the scale length in the longitudinal (flow) direction, r_h is the scale length in the lateral direction, and u_0 is the reference velocity amplitude of the oscillating velocity. This is the dimensionless form of the general equation of heat transfer for an inviscid ideal gas. Both the oscillating pressure and a temperature gradient are present in this equation. In regenerators and stacks used in thermoacoustic machines,[15] $c_p T_m \gg \omega u_m L$, therefore the coefficient of the third and fourth terms are both much smaller than the coefficient of the second term which is unity. This implies that the heat transfer due to oscillating pressure is small compared to heat transfer due to the temperature gradient. Ignoring those two terms and assuming the temperature gradient is dominated by mean temperature gradient in the x direction only, the heat transfer equation for oscillating flow, Eq. (19), was reduced to

$$\rho c_p \left(\frac{\partial T}{\partial t} + u \frac{\partial T_m}{\partial x} \right) = \nabla \cdot (k \nabla T). \tag{34}$$

After the spatial average and defining another heat transfer coefficient, h_T, as the heat transfer coefficient for oscillating flow subject to a temperature gradient, we obtain

$$\left\langle \rho c_p \frac{\partial T_1}{\partial t} \right\rangle + \rho c_p u \frac{dT_m}{dx} = -\frac{h_T}{r_h} \langle T_1 \rangle. \tag{35}$$

Solving for $\langle T_1 \rangle$, then equating $\langle T_1 \rangle$ with solution given in Eq. (12) by setting $p_1 = 0$, the heat transfer coefficient h_T can be expressed in terms of thermoviscous f functions

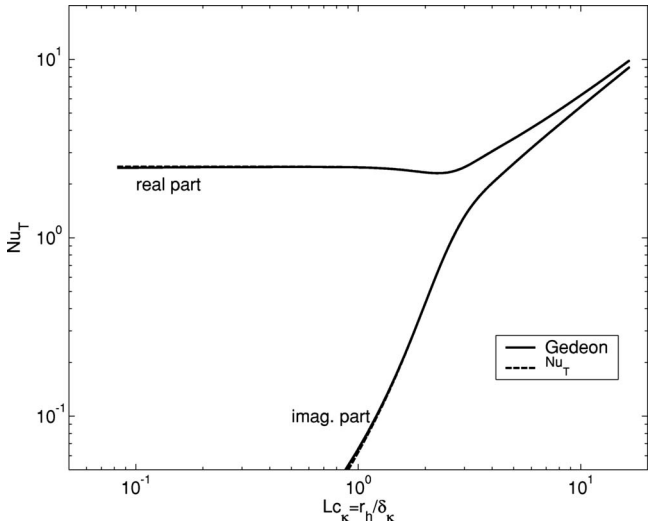


FIG. 4. Comparison between Gedeon's Nu_T and the results derived in this article. Only one line is seen for the real part of Nu_T because both results are in exact agreement.

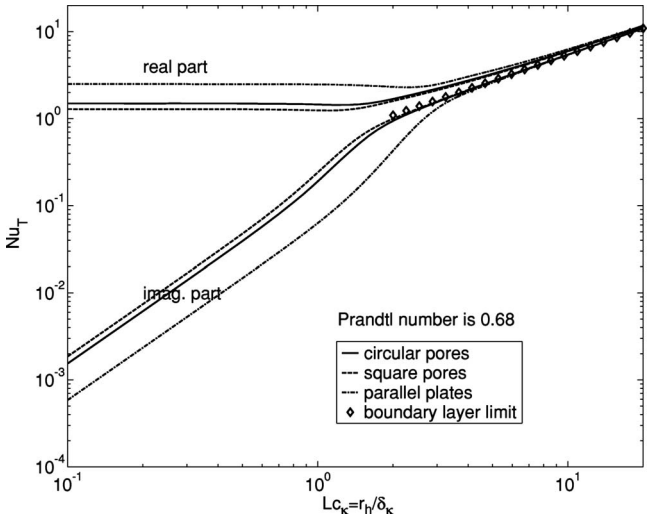


FIG. 5. Complex Nusselt number for circular pores, square pores, and parallel plates subject to a temperature gradient. The black diamond symbols are the boundary layer limit of Nu_T from Eq. (41).

$$h_T = -i\omega\rho_m c_p r_h\left[1 - \frac{(1-f_\nu)(1-\sigma)}{(1-f_\kappa)-\sigma(1-f_\nu)}\right]. \tag{36}$$

The complex Nusselt number for oscillating flow subject to a temperature gradient, Nu_T, is

$$\mathrm{Nu}_T = \frac{h_T r_h}{k} = 2i(Lc_\kappa)^2\left[\frac{(1-f_\nu)(1-\sigma)}{(1-f_\kappa)-\sigma(1-f_\nu)} - 1\right]. \tag{37}$$

Nu_p depends on thermal boundary layer thickness only, but Nu_T depends on both viscous and thermal boundary layer. That is because the temperature profile is related to the velocity profile when a velocity is also imposed. Gedeon[12] solved the mean-parameter model of Nu_T for oscillating flow between parallel plate. Gedeon's results and Nu_T from Eq. (37) are plotted in Fig. 4. The agreement is excellent for the real parts and for the imaginary parts at high Lautrec numbers. In Gedeon's paper, the imaginary part of Nu_T smaller than 0.1 was not plotted.

From Gedeon's paper,[12] using hydraulic diameter, $D_h = 4r_h$, as the characteristic length,

$$\lim_{\omega\to\infty} \mathrm{Nu}_T^{D_h} = 10 = \mathrm{Nu}_0. \tag{38}$$

Nu_0 is the Nusselt number for steady laminar flow between parallel plates under condition of uniform heat flux and when the film coefficient is in terms of section-average temperature rather than bulk (velocity-weighted section averaged) temperature.[16] Using r_h in this work, it converges to 2.5, as shown in Figs. 4 and 5. Therefore, when penetration depths are much larger than pores size, the steady-state flow results are applicable to oscillating flow.

Substituting Eqs. (27) and (28) into Eq. (36), using the fact that, $\delta_\nu = \sqrt{\sigma}\delta_\kappa$, and taking the limit $\delta_\kappa/r_h \to 0$, the boundary layer limit of h_T becomes

$$\lim_{\omega\to\infty} h_T = \frac{k}{\delta_\kappa}(1+i)\frac{1-\sqrt{\sigma}}{1-\sigma}. \tag{39}$$

The coefficient, $C_T = (1-\sqrt{\sigma})/(1-\sigma)$ is Prandtl number dependent and is summarized in Table I. For $\sigma = 0.68$, Eq. (39) becomes

$$\lim_{\omega\to\infty} h_T = 0.55\frac{k}{\delta_\kappa}(1+i). \tag{40}$$

When the thermal penetration depth is smaller than the hydraulic radius, the heat transfer coefficient is proportional to the thermal conductivity of the gas divided by the thermal penetration depth. Mozurkewich approximated the coefficient to be 0.61 for $\sigma = 0.68$ from the eigensolutions of a boundary-value problem he formulated to study the time-averaged temperature distribution in a thermoacoustic stack.[17] The gas temperature he used was that at the center of the pore. He also pointed out that this coefficient is Prandtl number dependent.

From Eq. (39), the boundary layer limit of Nu_T becomes

$$\lim_{\omega\to\infty} \mathrm{Nu}_T = \frac{r_h}{\delta_\kappa}(1+i)\frac{1-\sqrt{\sigma}}{1-\sigma}. \tag{41}$$

The boundary layer limit is plotted in Fig. 5 in black diamond symbols. Figure 5 shows the complex Nusselt number for oscillating flow inside circular pores, square pores, and parallel plates subject to a temperature gradient and choosing the Prandtl number to be 0.68.

The boundary value limit for h_T is important for the design of heat exchangers used in thermoacoustic machines. In addition to the value of C_T calculated here (0.55) and

TABLE I. C_T: the boundary layer limit coefficient of h_T as a function of the Prandtl number.

Prandtl number (n)	0.1	0.68	1.5	2	2.5	3	3.5
C_T	0.76	0.55	0.45	0.41	0.39	0.37	0.35

calculated by Mozurkewich[17] (0.61), Garrett *et al.*[18] assumed $C_T=1$, but then reduced its value to 0.76 to account for the sinusoidal velocity variation. Wakeland and Keolian[19] gave a value of 0.45 at small gas displacements, although a value of unity is approached when gas displacements are comparable to the thickness of the heat exchanger which is frequently the design target for heat exchangers used in thermoacoustic machines. The specific values of the boundary layer limit depend on which assumptions and what kind of averaged temperature are chosen. In this study the spatially averaged temperature is used, whereas Mozurkewich used the temperature at the center of the pore and it is also an approximation using the first term of the eigen-value summation only. Wakeland averaged the temperatures for pairs of hot and cold heat exchangers.

Application of similitude at $\mathrm{Lc}_\kappa > 1$ guarantees that

$$\lim_{\omega\to\infty} h \propto \frac{k}{\delta_\kappa}. \tag{42}$$

Because the velocity gradient and temperature gradient both depend on the penetration depth when $\mathrm{Lc}_\kappa > 1$, the heat transfer coefficient and the Nusselt number are geometry independent. However, when $\mathrm{Lc}_\kappa < 1$, both the velocity gradient and temperature gradient are characterized by the pore size so that the Nusselt number and heat transfer coefficient are geometry dependent. These features are shown in Figs. 3 and 5.

IV. CONCLUSIONS

It is shown that heat transfer in oscillatory flow requires a complex Nusselt number and the complex Nusselt number can be related to the thermoviscous f functions. The complex Nusselt number for heat transfer due to compression and expansion of oscillating flow, Nu_p, and the complex Nusselt number for oscillating flow subject to a temperature gradient, Nu_T, were calculated for the first time in terms of the thermoviscous f function. The agreement with Gedeon's results for Nu_T is excellent.

By relating the complex Nusselt number to the thermoviscous f function, the heat transfer behavior of oscillating flow in porous media can be characterized more quickly and accurately without measuring the temperature or requiring auxiliary heat exchangers. It has also been shown that the f-function interpretation of the Nusselt number for oscillating flow produces results that are consistent with several previous theoretical and experimental investigations of heat exchangers for oscillating flow in the appropriate limit: $\mathrm{Lc}_\kappa \geqslant 1$.

ACKNOWLEDGMENTS

The authors are grateful for the support of this investigation provided by the Office of Naval Research and the Educational and Foundational Research Program funded by the Penn State Applied Research Laboratory.

[1] R. Stirling, "Improvements for diminishing the consumption of fuel, and in particular an engine capable of being applied to the moving (of) machinery on a principle entirely new," English Pat. No. 4081 (1816).
[2] S. Backhaus and G. W. Swift, "A thermoacoustic Stirling-heat engine," Nature (London) **399**, 335–338 (1999); J. Acoust. Soc. Am. **107**, 3148–3166 (2000).
[3] S. Backhaus and G. W. Swift, "Fabrication and use of parallel plate regenerators in thermoacoustic engines," *Proceedings of the 36th Intersociety Energy Conversion Eng. Conference*, 29 July–2 Aug 2001; IECEC2001-CT-41.
[4] J. Liu and S. Garrett, "Characterization of regenerator materials for thermoacoustic refrigeration," J. Acoust. Soc. Am. **115**, 2380 (2004).
[5] F. M. White, *Mass and Heat Transfer* (Addison-Wesley, Reding, MA, 1991).
[6] D. Gedeon and J. G. Wood, "Oscillating-flow regenerator test rig: Hardware and theory with derived correlations for screens and felts," NASA Contractor Report 198442, NASA-Lewis Research Center, February, 1996.
[7] G. W. Swift, *Thermoacoustics: A Unifying Perspective for Some Engines and Refrigerators* (Acoustical Society of America, New York, 2002).
[8] G. W. Swift, "Thermoacoustic engines," J. Acoust. Soc. Am. **84**, 1145–1180 (1988).
[9] L. A. Wilen, "Measurements of thermoacoustic function for single pores," J. Acoust. Soc. Am. **103**, 1406–1412 (1998).
[10] A. A. Kornhauser and J. L. Smith, Jr., "Heat transfer with oscillating pressure and oscillating flow," 24th IECEC 1989, pp. 2347–2353.
[11] A. A. Kornhauser and J. L. Smith, Jr., "Application of a complex Nusselt number to heat transfer during compression and expansion," J. Heat Transfer **116**, 536–542 (1994).
[12] D. Gedeon, "Mean-parameter modeling of oscillating flow," J. Heat Transfer **108**, 513–518 (1986).
[13] L. D. Laudau and E. M. Lifshitz, *Fluid Mechanics* (Pergamon, New York, 1959).
[14] G. W. Swift and W. C. Ward, "Simple harmonic analysis of regenerators," J. Thermophys. Heat Transfer **10**, 652–662 (1996).
[15] S. L. Garrett, "Resource Letter TA-1: Thermoacoustic engines and refrigerators," Am. J. Phys. **72**, 288–293 (2004).
[16] M. Jacob, *Heat Transfer* (Wiley, New York, 1949).
[17] G. Mozurkewich, "Time-average temperature distribution in a thermoacoustic stack," J. Acoust. Soc. Am. **103**, 380–388 (1998).
[18] S. L. Garrett, D. K. Perkins, and A. Gopinath, "Thermoacoustic refrigerator heat exchangers: Design, analysis and fabrication," *Proceedings of the Tenth International Heat Transfer Conference*, 1994, Vol. **4**, pp. 375–380.
[19] R. S. Wakeland and R. M. Keolian, "Effectiveness of parallel-plate heat exchangers in thermoacoustic devices," J. Acoust. Soc. Am. **115**, 2873–2886 (2004).

A study of the vibro-acoustic modulation technique for the detection of cracks in metals

Philippe Duffour
Department of Civil and Environmental Engineering, University College London, London, United Kingdom

Marco Morbidini and Peter Cawley[a)]
Mechanical Engineering Department, RCNDE (UK Research Centre for NDE), Imperial College London, London, United Kingdom

(Received 15 April 2005; revised 4 October 2005; accepted 29 November 2005)

One implementation of the vibro-modulation technique involves monitoring the amplitude modulation of an ultrasonic vibration field transmitted through a cracked specimen undergoing an additional low frequency structural vibration. If the specimen is undamaged and appropriately supported, the two vibration fields do not interact. This phenomenon could be used as the basis for a nondestructive testing technique. In this paper, the sensitivity of the technique is investigated systematically on a set of mild steel beams with cracks of different sizes and shapes. A damage index was measured for each crack. The correlation obtained between the crack size and the strength of the modulation is fairly poor. The technique proved extremely sensitive to the initial state of opening and closing of the crack and to the setup due to the modulating effects of contacts between the specimens and the supports. A simple model is proposed which explains the main features observed and approximately predicts the level of sideband obtained experimentally. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2161429]

PACS number(s): 43.35.Zc [YHB] Pages: 1463–1475

I. INTRODUCTION

Acoustic nonlinear phenomena applied to nondestructive evaluation have been an area of intense study since the 1960s.[1] The remarkable strength of many nonlinear effects introduced by defects makes their use for damage detection very attractive. Two main nonlinear effects have been studied extensively for NDT applications: (1) measuring the wave speed dependence on the wave amplitude;[2] (2) monitoring the generation of harmonics.[3] In both cases the evolution of the nonlinear effect can be studied over time (e.g., for fatigue monitoring) or in space for defect location.[4] Contact acoustic nonlinearities (CANs) are among the most efficient mechanisms generating nonlinear features.[5] CANs comprise all solid-to-solid interfaces with imperfect adhesion, such as cracks, disbonds, etc. Vibro-acoustic modulation is based on this kind of nonlinearity.

The phenomenon was first reported in a 1975 US patent[6] but the bulk of the scientific literature on the subject arose in the mid-1990s.[7,8] It can be described as follows: if an undamaged specimen is subjected to both a low frequency vibration *and* an ultrasound probing signal, the resultant field is simply the superposition of the fields obtained by applying the two excitations separately. However if the system is damaged (e.g., by a crack or some debonding), the two vibrations interact so that the ultrasound wave is modulated in amplitude and/or phase[9] by the low frequency oscillation. Therefore, the occurrence of amplitude modulation in such a setup can be taken as an indication of the presence of damage in the specimen and could be used as the basis of a nondestructive technique for damage detection. In practice, the phenomenon is usually studied in the frequency domain where the modulating effect is manifested by the presence of sidebands around the main peak at the ultrasonic frequency.

Using Vibro-Acoustic Modulation (VAM), Donskoy *et al.*[10] successfully detected cracks in water pipes, measuring sidebands 30 dB higher compared to a reference measurement on a pipe without cracks. The detection was performed using one single ultrasonic frequency and the amplitude of the sidebands was shown to be proportional to the amplitude of the transmitted probing ultrasonic signal. Ekimov *et al.*[11] used torsional ultrasonic waves at one frequency to detect an artificially introduced interfacial defect. A thin aluminum plate with a single fatigue crack was studied by Zaitsev *et al.*;[12] the amplitude of the modulation sidebands was demonstrated to be dependent on the amplitudes of the high and low frequency vibrations, and on the global Q-factor of the system, the Q-factor being modified by attaching strips of insulating tape to the test structure. Several other examples of successful applications of VAM to NDT testing were later reported in Plexiglas™, sandstone, and in automotive engines.[13] The sizing capabilities of the VAM testing were first experimentally studied by Donskoy *et al.*[14] using a set of laboratory specimens containing calibrated cracks. Here the dependence of the sensitivity of the technique on the chosen ultrasound frequency was taken into account by averaging the amplitude of the sidebands over nine measurements at nine ultrasonic frequencies between 110 and 150 kHz with 5-kHz increments. A damage index was defined by this frequency averaging, after having normalized the amplitude of the sidebands with respect to the amplitude of both the ultra-

[a)]Author to whom all correspondence should be addressed; electronic mail: p.cawley@imperial.ac.uk

sound and low frequency vibrations. From the modeling point of view, the phenomenon was usually explained in terms of increasingly complex nonlinear models of the wave/crack interaction. The crack was typically modeled as a nonlinear spring.[9,15] When the ultrasound frequency is fairly low, the system involves ultrasound standing wave fields,[7,10] whereas for higher ultrasound frequencies, traveling waves are a more suitable description.[9,16] The interaction between surface waves and cracks in the presence of residual stresses around the crack was studied in detail by Kim *et al.*[16,17]

The objective of this paper is to provide a systematic assessment of the potential of a particular implementation of the VAM method for NDT of metallic structures where small fatigue cracks may be present. It is of particular interest to identify quantitatively the sensitivity of the method to crack size, and how well it performs with respect to sizing. The minimum level of low frequency strain necessary for it to work also has to be determined for the method to be robust in a real inspection setting. With this objective in mind, the best choice of ultrasonic frequencies, and finally the susceptibility of the technique to noise must also be identified. A simple model of the phenomenon is proposed to provide some insight into the main features observed and to explain semiquantitatively the trends in the experimental results.

II. EXPERIMENT

In this study, the sensitivity of the proposed VAM technique to the presence of a crack is measured by the ratio of the average of the first left and first right sideband amplitudes over the amplitude of the main ultrasonic peak. The effect of changing the ultrasonic frequency on the sideband amplitudes was experimentally investigated over a broad ultrasonic frequency range (50–230 kHz). By doing this, the ultrasonic frequencies that give maximum sideband amplitudes could be found and their distribution in the frequency domain analyzed. The dependence of the sideband-carrier ratio curves on the level of strain generated by the low frequency vibrations at the crack was also measured. The study was carried out using a laboratory set of cracked mild steel beams with one through width crack for each beam. These cracks were obtained by fatigue and propagated to different depths. In this way a damage index was estimated for each crack, and its value compared to the actual cracked section estimated after the beams were broken in half. A first implementation of VAM testing was realized exciting the first bending resonance of the beams by means of a shaker; later an impact hammer was used. It was then possible to compare the results obtained using these two different testing rigs so that their reliability for performing the crack detection can be discussed.

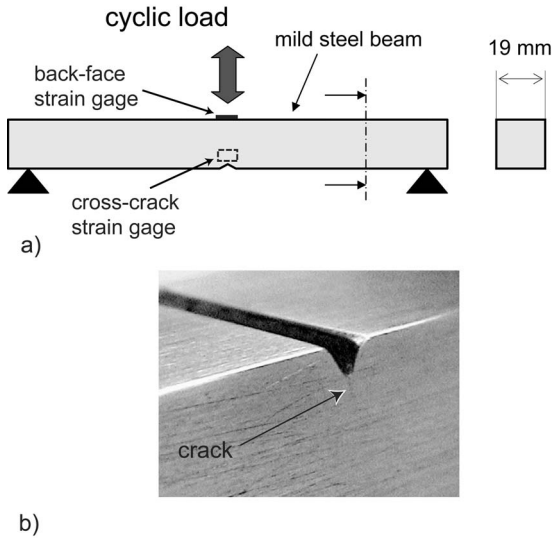


FIG. 1. (a) Schematic of the fatigue rig setup. The same setup was later used for estimating the initial partial opening of the crack of specimen "CO" by applying one back face and one cross-crack strain gauge. (b) Photograph of the crack growing from the root of the notch.

A. Description of the cracked laboratory specimens

A set of seven mild steel beams were cracked to be tested in the experiments described in this paper. A square cross section having a nominal side dimension of 3/4 in. (approximately 19 mm) and lengths of 200, 250, and 300 mm (named, respectively, S, M, and L) were chosen for these test specimens. The material used was mild steel for easier machining, as specified by standards BS970 and EN3B (AISI C1020 for US equivalent). Each beam was "V" notched at midspan to a depth of about 2 mm, and then fatigued at a rate of 10 cycles/s, in a three-point bending rig, as shown in Fig. 1. The fatigue settings used are given in Table I for each specimen. "Load" is the maximum compression load reached at every fatigue cycle and "N cycle" is the number of cycles at the termination of the fatiguing. One crack for each specimen was produced in this way. The crack developed through the thickness of the beams after it initially

TABLE I. Details of fatigue settings and crack sizes.

Beam	L1	M1	S1	L2	M2	S2	CO
Length (mm)	300	250	200	300	250	200	250
Load (kN)	6	6.5	7	7	8.3	8.5	7
Number of cycles ($\times 10^3$)	137	38.2	148	33.4	34.7	29.2	52.1
Edge 1 (mm)	2.9	2.2	2.5	3.1	1.7	1.6	1.6
Edge 2 (mm)	1.4	1.3	2.3	2.3	1.6	1.6	1.1
Max depth (mm)	3.4	2.4	3.6	3.4	2.3	2.1	1.8
% cracked area	15.5	11.1	16.6	16.1	10.6	10.4	8.5
Roughness (μm)	7.9	8.9	6.5	16	7.6	9.6	8.4
Mean peak spacing (μm)	183	230	199	353	171	218	202

formed at the root of the notch, and its propagation was monitored using a microscope. In this way the growth of the crack could be stopped at different depths. In order to obtain an accurate definition of the profile of each crack, the samples were later broken in half, and the crack depth at the edges and at the point of maximum depth (the crack profiles all exhibited a slight concavity) were measured by taking pictures of the crack profiles using a magnification factor of about 10. The roughness of the crack surface and the average spacing between profile peaks were measured using a Talysurf surface profiling machine, scanning along a line across the width of the crack from one edge to the other. These measurements are reported in Table I.

Using this procedure, two groups of cracked samples (labeled 1 and 2) were produced from two different material batches. The specimens L1 and S1 were cracked using a higher number of cycles compared to the specimen M1; this could result in a different morphology of the cracks. However, the specimens L2, M2, and S2 were all cracked using similar loads and number of cycles so that it is likely that the respective cracks differ only because of their size and not because of their morphology.

An extra specimen labeled "CO" (for Crack Opening) was produced later to investigate the initial opening of the cracks produced by this setup. The steel used for this specimen was the same as that for specimens S2, M2, and L2 and the fatigue settings for the production of the crack were also similar. The results for this beam will be reported later in the paper.

B. Results using a continuous harmonic excitation rig

Two practical implementations of the VAM technique were investigated in this study. First, the cracked beams were supported horizontally on two foam pads and instrumented with two permanently bonded PZT discs of diameter 10 mm and thickness 4 mm for the transmission and measurement of the ultrasound typically in the range 50–230 kHz. The input PZT disc was driven by an amplified sine burst of 30 kcycles while the ultrasonic (UT) and strain output time traces were logged using a storage oscilloscope after appropriate filtering. The data were stored and processed later using standard FFT tools. An electromagnetic shaker (Data Physics model S-004) was attached to the beams in the proximity of the notch and in the direction perpendicular to its length; the connection was realized by first bonding an aluminum stud with a threaded hole to the beams and connecting this to the shaker via a push rod. The frequency of the force applied to the beams was tuned to the resonance of the first bending mode of the beams at approximately 1 kHz so as to maximize the amplitude of the vibration and thus the "crack breathing." The amplitude of the low frequency vibration was varied by changing the output of the shaker amplifier (Data Physics model A-060). The resulting opening and closing of the crack was indirectly measured by the backface strain gauge of Fig. 1 (the cross-crack strain gauge was not fitted in these tests). With this setup, it was possible to automate the frequency sweep of the ultrasound over broad frequency ranges, while the low vibration was kept at a constant amplitude. This procedure was then repeated for increasing levels of the amplitude of the low frequency vibration up to a maximum of 80 micro-strain (peak value) measured at the strain gauge.

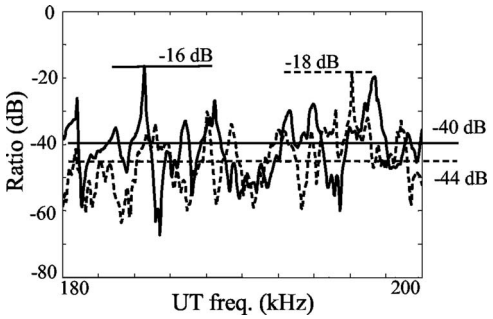


FIG. 2. Comparison between the ratio of the amplitudes of the average first sidebands over the ultrasonic fundamental peak, at 40 microstrain, for one cracked (solid line) and one uncracked (dashed line) specimen. The maximum ratio is 2 dB higher for the damaged beam whereas the average ratio over the frequency range 180–200 kHz is 4 dB higher.

The same experiment was repeated using uncracked specimens in order to (a) verify that the amplitude modulation of the ultrasound, and hence the sidebands, are only caused by the crack and (b) establish appropriate criteria for the comparison between damaged and undamaged specimens. Typical results from this comparison are shown in Fig. 2 where data from both a cracked and an uncracked specimen are plotted for the same level of strain (40 microstrain peak value measured at the middle cross section). Here the sideband to carrier ratio is plotted against the ultrasonic frequency. As can be seen from Fig. 2 no clear indication of damage could be obtained, either by using the maximum ratio criterion as an indicator of crack presence or from the value of the average ratio over the whole frequency range considered. In the former case a difference of just 2 dB was measured, and only a slight improvement (4 dB) was obtained in the latter.

It is believed that the poor reliability obtained with this rig is due to the contact nonlinearities between the shaker attachment and the beam, as well as the beam and its supports, causing an amplitude modulation of the ultrasound which resembled that generated by the crack (the nature of this interaction would require further investigation). Various rubber washers or pads were inserted between the beam and the shaker attachment as an attempt to reduce this contact nonlinearity but this proved unsuccessful as the strain excited in the beam is then reduced drastically.

C. Results using the impact hammer suspension (IHS) rig

A suspension rig was set up to overcome the reliability problems of the shaker rig, and is shown schematically in Fig. 3. The suspension was realized by bonding two rubber studs at the nodal position of the first bending mode of the beam and then by passing two elastic strings through them. The beam was instrumented with two PZT discs and with a strain gauge of the same kind described earlier. The excitation of low frequency modes was performed by hitting the beam at about midspan using an automatic tapping device.

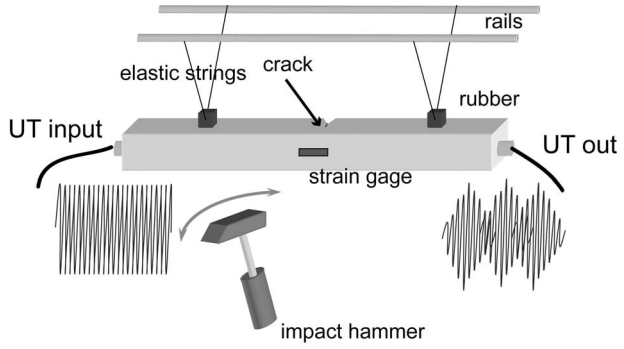


FIG. 3. Schematic of the impact hammer suspension (IHS) rig.

The tip of the hammer was made of a soft nylon material so that the dominant mode excited was the first bending mode, at around 1 kHz. Having automated the tapping, it was possible to automate the measurement routine. After every hit the ultrasonic frequency was incremented by 30 Hz and the measurements repeated. In this way the ultrasonic frequency could be swept automatically over a large frequency range (50–230 kHz). Another major advantage of this rig was that the amplitude of the sidebands for decreasing levels of the strain could be analyzed along the transient decay of the first bending mode following every impact. A comparison between a cracked and uncracked specimen is shown in Fig. 4 within a restricted frequency range. As expected, the amplitude of the ultrasonic carrier (group of overlapping curves in the upper part of the plots) is virtually independent of the amplitude of the low frequency bending strain. The ratio of the average of the amplitudes of the first sidebands over the carrier is clearly strain dependent only in the presence of a crack. For the uncracked specimen, the ratio curves (obtained with the same postprocessing routines) tend to a common level at about −65 dB, with two peaks at about −50 dB at the antiresonances of the ultrasonic carrier. This might be due to the small amplitude of the carrier signal at its antiresonances, causing errors in the ratio value. However, if a crack was not present, the high and low frequency vibration fields did not interact significantly and hence the IHS rig was suitable for further testing. These results also confirmed that the shaker attachment and the supports did produce the nonlinearity seen in the shaker rig experiments on uncracked beams. Using the IHS rig, in the case of the cracked specimen L2, the maximum ratio was about −25 dB at around 80 microstrain rms, allowing a good dynamic range for performing the

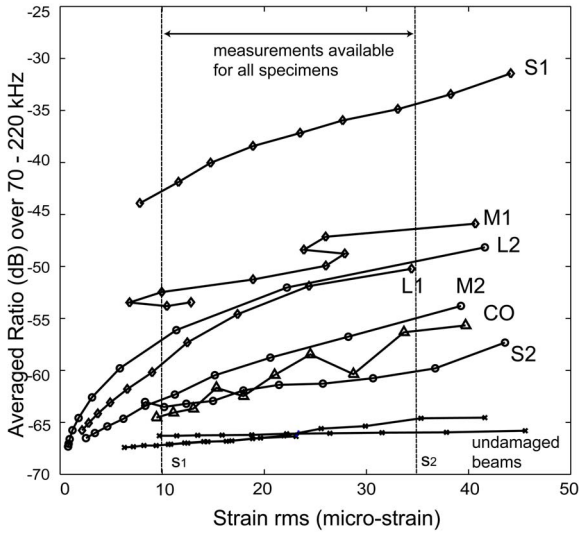


FIG. 5. Results of the VAM tests using the IHS setup for the six cracked beams and for three uncracked beams. The irregular behavior of the specimen M1, with respect to the low frequency vibration, is due to beating of the first flexural mode about the two principal axes. This was absent for all the other beams.

crack detection by comparing the maximum values of the ratio across a chosen frequency range. However this dynamic range could be further improved by considering the average ratio over the whole frequency range. It is also interesting that local maxima of sensitivity seem to be present every 3 to 5 kHz, starting at about 100 kHz. This suggests that a sweep over about 5 kHz would be sufficient to ensure a high sensitivity point.

Definition of a damage index. The experiments were repeated for all the six cracked beams available. For each beam the value of the average ratio (i.e., the average sensitivity) $\bar{R}(s)$ across the whole frequency range swept from 70 to 220 kHz was plotted as a function of the strain level of the low frequency vibration. Then the strain range within which experimental points were available for all the specimens was considered, from s_1 to s_2 (see Fig. 5). A damage index (DI) was derived from the area of the graph delimited by the sensitivity versus strain plot (on a linear scale) for each damaged specimen, and that under the ratio versus strain curves for the undamaged specimens of the same size,

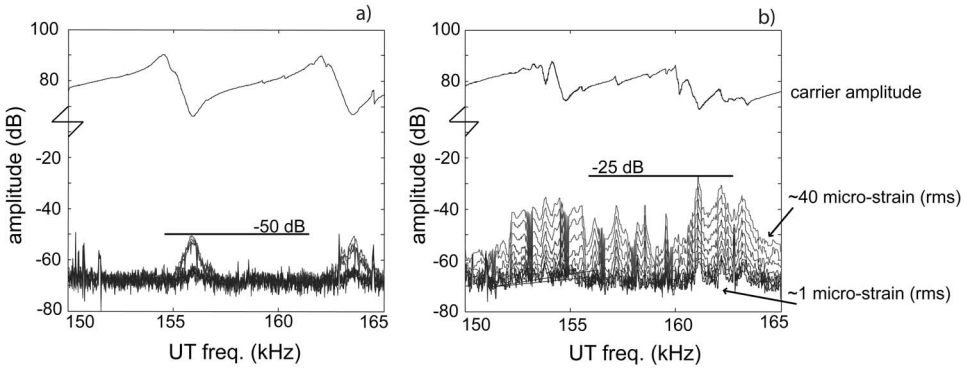


FIG. 4. Plot of the carrier amplitude (top curve) and first average sideband amplitudes (bottom curves) at different strain levels for carrier frequencies ranging from 150 to 165 kHz: (a) uncracked beam, (b) cracked beam.

$$\mathrm{DI} = \frac{\int_{s_1}^{s_2} \overline{R_{\mathrm{cracked}}}(s)\,ds}{\int_{s_1}^{s_2} \overline{R_{\mathrm{uncracked}}}(s)\,ds}. \quad (1)$$

The damage index defined in Eq. (1) takes into account the dependence of the sensitivity of the VAM technique on the ultrasonic frequency and on the level of the low frequency strain reached during the tests. It also takes into account the nonlinearities that may be introduced by the supports and attachments by dividing the integral for the damaged specimen by the same quantity for undamaged specimens. The DI will therefore be a value greater than 1 for cracked specimens. If the same specimens were tested using different implementations of the VAM test, reaching similar levels of the low frequency vibration and sweeping the ultrasound across the same frequency range, the DI as defined in Eq. (1) would also quantify the reliability of the different implementations. Results using this damage index will be shown in Sec. IV.

III. SIMPLIFIED MODELING

The following presents a simple model of the vibro-modulation phenomenon. The purpose of the model is to provide an explanation for the main features seen in the experiments and to investigate whether a simple approach can yield order of magnitude prediction of the sideband amplitudes observed.

A. General principle of the model

The geometry of the beam modeled was that of the long beams used in the experimental study. The dimensions were $300 \times 20 \times 20$ mm and the material properties used were those of mild steel. Their first bending mode frequency was around 1 kHz while the ultrasound frequency used was around 200 kHz. Therefore for every cycle of low frequency vibration, 200 ultrasonic cycles are transmitted. It will be assumed throughout the modeling section that in these conditions, the UT vibration is in steady state within each half bending cycle. To verify this assumption, the UT Q-factors were measured experimentally on some of the specimens. Q-values up to 130 kHz were measured by the half-power bandwidth method from the carrier amplitude versus frequency curves, an example of which was shown in Fig. 4. The measured Q-factors were mostly around 200. Given this damping value, a vibration at 200 kHz reaches approximately 80% of its steady-state value during each half of the bending period so the assumption is not entirely valid but comes close to being met.

Results are presented for different values of Q; it will be appreciated that the prediction will be inaccurate at Q above about 200. However, the general trends of the modulation at different frequencies and strains can still be obtained.

In these conditions an adequate way of describing the relationship between the ultrasonic input and output is to introduce the transfer function between the two points where the sensors are located. Transfer function measurements[18] alone have been used for crack detection for some time and several papers model the effect of cracks on transfer functions[19,20] but the resonance frequency shifts introduced by cracks often proved not to be sensitive enough for many practical applications. By contrast, in a vibro-modulation setup, two types of vibration interact. As the beam bends according to its first mode shape, the degree of opening and closing of the crack changes so that the transmission characteristics of the cracked cross-section change. The transfer function then gradually shifts from the state when the crack is most open to the state where it is the most closed. This is illustrated diagrammatically in Fig. 6.

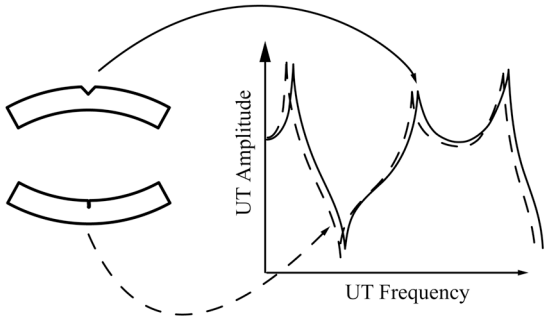


FIG. 6. Diagram showing how the transfer function from one end of the beam to the other shifts as the bending of the beam opens and closes the crack.

The same underlying idea was used by Kim *et al.*[16,17] In both their model and this one, the presence of sidebands is accounted for by the parametric geometrical deformation of the crack induced by the low frequency vibration. However the operational conditions of the system they investigated were quite different. In the case of Kim *et al.*, this geometrical deformation was probed *locally* using a SAW at a frequency of around 1 MHz in a pulse-echo configuration. In the model described here, the geometrical deformation of the crack affects the 200-kHz ultrasound field *globally* by slightly shifting the beam resonances over the course of a bending cycle.

A pure sine wave at frequency f_0 and modulated in amplitude in the time domain results in the appearance of smaller peaks on either side the main peak centered on f_0 in the frequency domain. As before, the term "carrier" will refer to the pure sine wave, i.e., ultrasound, and the first smaller peaks on both sides of the carrier are called "sidebands."

It is assumed here that a closed crack is transparent to the ultrasound. Therefore the transfer functions in the closed state were calculated for beams without crack. This amounts to neglecting the wave scattering properties of closed cracks.[21] However at 200 kHz, the wavelength in steel is sufficiently large for this simplification to be justified. The transfer function in the open state was obtained by considering the crack fully open for various crack depths.

This type of modeling bears strong similarities to an earlier study by Richardson,[22] who investigated the propagation of a wave through a planar interface with no tensile strength. However, in his one-dimensional model, the two sides of the clapping interface could only move parallel to each other so that the contact could either be complete or

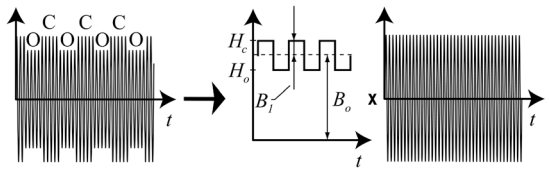


FIG. 7. Diagram detailing the modulating effect when the cracked region is free of compressive residual stresses. The modulated signal can be decomposed into a pure high frequency sine wave multiplied by a square modulating function. B_0 is the mean value of the modulating signal. B_1 is the peak amplitude variation around this mean. $H_{o/c}$ is the amplitude of the transfer function at a given particular ultrasonic frequency when the crack is *open/closed*.

nonexistent. Here, the low frequency bending stress influences the ultrasound by cyclically changing the area of contact between the crack faces. This is different from some of the work on vibro-modulation cited in Sec. I where the crack is usually modeled as a material nonlinearity having an elastic stiffness comprising a constant (linear) component plus a quadratic corrective term which varies with the low frequency stress.

It is also assumed here that the ultrasonic vibration—because of its higher frequency and much lower amplitude compared to the bending vibration—does not affect the opening and closing of the crack. For an ultrasound excitation at a given frequency f_0, the value of the transfer function at f_0 will then oscillate between two extreme values as the crack opens and closes (see Fig. 6). Exactly how the ultrasound frequency changes from one extreme value to the other depends on the particular characteristics of the crack. In particular, the presence of residual stresses in the cracked region is likely to affect the degree of opening and closing of the crack induced by bending. Pursuing the analysis further requires making simplifying assumptions. First the case of a crack free of residual stresses is investigated. Then a simple model will be used to study how compressive stresses in the beam and initial crack opening interact with the cyclic crack closure caused by bending.

B. Case with no residual stresses

1. Formulation of the problem

Imagine that when the beam is at rest, the crack faces are two mating surfaces in "grazing contact" with no residual compressive stresses. Then the crack immediately and fully opens (closes) as soon as the bending stresses in the top half of beam become tensile (compressive), assuming the crack is at the top half of the beam. In this idealized situation, the magnitude of the bending strain has no influence on the crack closure. For half the bending cycle the ultrasound settles in a system with a completely open crack whereas during the other half of the cycle, the crack is closed. The measured ultrasonic output will then be a high frequency sine wave modulated in amplitude by a low frequency square wave as illustrated in Fig. 7.

This modulated signal can be viewed as a pure sinusoidal oscillation multiplied in the time domain by a square signal. This simple multiplication becomes a convolution of the corresponding spectra in the frequency domain.[23] At a given ultrasonic frequency f_0, the modulating square signal oscillates between the values of the two transfer function values $H_o(f_0)$ (open) and $H_c(f_0)$ (closed). This modulation function must have a nonzero dc value to account for the fact that ultrasound is transmitted even when the crack is fully open. This dc value is simply the mean of H_o and H_c which can be written, using the notation defined in Fig. 7,

$$B_0 = \frac{H_o + H_c}{2}. \tag{2}$$

Around the mean, the amplitude of the square oscillation is

$$B_1 = \frac{H_c - H_o}{2}. \tag{3}$$

If X_m denotes the Fourier transform of the modulating function, it can be written in the form of a series as follows:

$$X_m = \frac{H_o + H_c}{2}\delta_0 + (H_c - H_o)\left(\frac{1}{\pi}\delta_{f_m} + \frac{1}{3\pi}\delta_{3f_m} \ldots\right), \tag{4}$$

where δ_{kf_m} is the delta function at frequency $k.f_m$ with k an integer. Using the coefficients B_0 and B_1 defined in Fig. 7, Eq. (4) can be more simply rewritten:

$$X_m = B_0\delta_0 + B_1\left(\frac{2}{\pi}\delta_{f_m} + \frac{2}{3\pi}\delta_{3f_m} \ldots\right). \tag{5}$$

When this modulating function is convolved with a pure sine wave of amplitude A and frequency f_0, $X_0 = A\delta_{f_0}$, the first dominant terms of the convolved spectrum are

$$X_0 * X_m = A\frac{H_o + H_c}{2}\delta_{f_0} + A\frac{(H_c - H_o)}{\pi}(\delta_{f_0+f_m} + \delta_{f_0-f_m}) + \ldots$$

$$= \underbrace{AB_0\delta_{f_0}}_{\text{Carrier}} + \underbrace{AB_1\frac{2}{\pi}(\delta_{f_0+f_m} + \delta_{f_0-f_m})}_{\text{First sidebands}} + \ldots . \tag{6}$$

Therefore the amplitude ratio, R, between first sideband amplitude and the carrier of the convolved spectrum is

$$R = \frac{2}{\pi}\frac{|H_o(f_0) - H_c(f_0)|}{|H_o(f_0) + H_c(f_0)|} = \frac{2}{\pi}\frac{B_1}{B_0}. \tag{7}$$

This ratio quantifies the parametric effect introduced by the crack. If there is no crack, there is no difference between H_o and H_c and the ratio is zero. If the crack is large, one can expect the shift between $H_o(f_0)$ and $H_c(f_0)$ also to be large at most ultrasonic frequencies f_0; consequently the value of the ratio is also large. The coefficients B_1 and B_o only depend on the values of the transfer functions in the open and closed states. To obtain quantitative predictions for this ratio, it is therefore necessary to have plausible values for the two transfer functions $H_{o/c}(f)$, calculated for the beams considered in this study. Two methods were used for this purpose: first a receptance analysis, then a finite element (FE) model of the beam.

2. Simulation using receptance analysis

Receptance analysis[24] is a technique that is particularly useful when it is necessary to predict the vibration response of a system which can be decomposed into simpler sub-

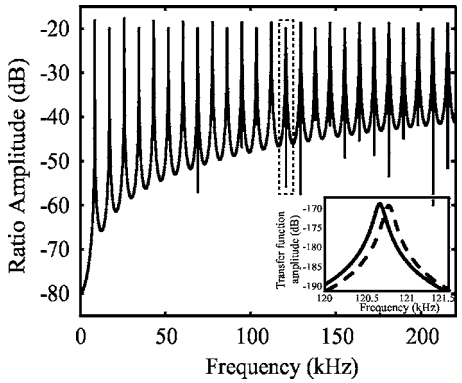


FIG. 8. Ratio of the sideband amplitude over carrier amplitude for varying ultrasound carrier frequency for an equivalent 10% crack depth and $Q=1000$. Inserted is a plot of the corresponding transfer function amplitudes, when the crack is closed (dashed) and open (solid) in the frequency range of the boxed peak on the main plot.

systems, for which transfer functions are easily obtained. The response of the initial complex system can then be computed by coupling these subsystems at appropriate points. This technique was used here to model very crudely the cracked beams. The beams are decomposed into three separate plain uncracked beams: the cracked section was modeled by a very short beam (1 mm long) of reduced cross section compared to the two beams on the sides. These two side beams are identical, 150 mm long, so that the overall length is roughly 300 mm. They both have a 20×20 mm cross section. To simulate various defect sizes, the cross-section reduction of the short middle beam was varied from an initial 10% representing the notch up to 20% (percentage of side beam cross-section area). Note that the 10% area reduction representing the notch was added for completeness but since this reduction was included for all cases, it does not act as a defect: the notch, being always open, does not produce modulation so that the defect was effectively varied from 0 to 10%.

The transfer functions for each elementary beam were computed using the standard analytical formula:[25]

$$H_{ij}(\omega)=\sum_n \frac{u_i^n u_j^n}{\omega_n^2+i\omega_n\omega/Q_n-\omega^2}, \tag{8}$$

where $H_{ij}(\omega)$ denotes the value of the transfer function at the frequency ω between point i and point j. u_i^n is the mode shape value at location i for mode n, ω_n and Q_n are the natural frequency and the Q-factor for mode n. Only the longitudinal modes were included in this analysis. This is a one-dimensional model so no axial-bending coupling is possible. The natural frequencies and mode shape coefficients were determined analytically for each sub-beam using the standard theory for the longitudinal vibration of uniform bars.[25] All the longitudinal modes up to 1 MHz were included. At this stage in the analysis, the value of the Q-factor, Q_n in Eq. (8), was set to 1000 for all modes. The three beams were then assembled together using the coupling formulas for linear systems in series.[24] The result from this simulation is shown in Fig. 8.

The quantity plotted is the ratio of the first sideband amplitude over the ultrasonic carrier amplitude, denoted R in Eq. (7). As mentioned in the previous section, this ratio, plotted here against the carrier frequency, gives a measure of the parametric effect caused by the defect. This plot was obtained with a crack size represented by a cross-section reduction of 10%. From Fig. 8, several important observations can be made. First, there are very large variations in the ratio amplitude with frequency (note that the vertical scale is in dB). The ratio reaches a maximum value of about −17 dB at peaks equally spaced in frequency. At those peak frequencies, the sideband amplitude is about 10% of the carrier amplitude. For some of the peaks (e.g., at 121 kHz, the peak framed in Fig. 8), a sharp drop-out can be seen. This feature should in fact appear for all the peaks but the frequency resolution of the plot makes them appear randomly. To understand the presence of these drop-outs and more generally the structure of this plot, it is useful to return to the transfer functions in the two extreme states (crack closed/crack open) in a narrow frequency range around a resonance. This is shown as an insert in Fig. 8 where the dashed curve represents the amplitude of the transfer function when the crack is closed (equivalent to the uncracked case), while the solid line is the amplitude when the crack is open. The two curves intersect at about 120.7 kHz. At this point there is no difference in the transfer functions between the open and closed states so that the amplitude ratio becomes zero. In decibels, this becomes minus infinity, hence the drop-outs observed previously. This plot also shows that the maximum difference between the two curves occurs near resonances. This explains why the ratio plot in Fig. 8 featured equally spaced peaks: the extensional modes of a plain beam are simple multiples of a constant, they are therefore equally spaced and so are the ratio maxima. The fact that the ratio is maximum near system resonances has two important additional consequences. First, as observed in the experiment, the UT frequency will have to be swept over some range in order to lock onto some system resonance (in the ultrasound domain) where the ratio shows the best sensitivity. Second, the fact that the difference between the two curves is maximum around resonances also points to the importance of the (ultrasonic) damping for this technique: the lighter the damping, the larger the magnification factor, the higher the peak value and the larger the difference between the two states.

In conclusion, the receptance model provides an explanation for the strong dependence of the magnitude of the sidebands with the ultrasound frequency and damping. The reasons for this are clear when the modal density is low and peaks are well separated. However, the modal density at 200 kHz for the beams considered here will be considerably larger than it appeared in this model since flexural and other modes will also be excited. It remains to be seen whether the conclusions reached in this section hold for a higher modal density.

3. Simulation using a finite element model

To confirm the previous conclusions, the beam was modeled as a three-dimensional solid using the finite element package ABAQUS. The model was as close as possible to the specimen used in the experiment. The dimensions were $300\times20\times20$ mm, the notch was located at midlength, with

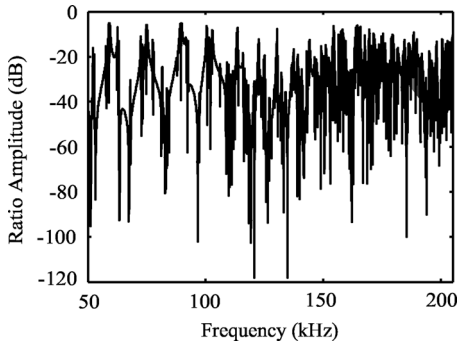


FIG. 9. Amplitude ratio of first sidebands over fundamental against fundamental frequency resulting from FE computation with 10% crack depth and $Q=1000$.

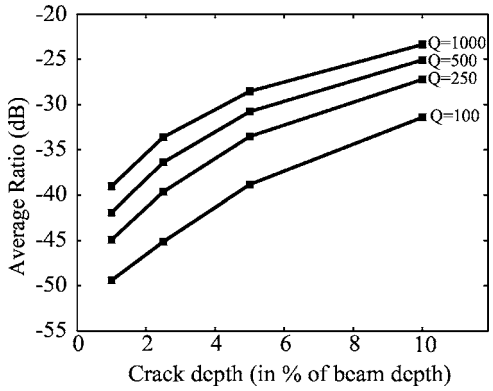


FIG. 10. Plot showing the amplitude ratio of sidebands over fundamental against the defect size, for various values of the Q-factor.

a 90° aperture and 2 mm depth. The crack was modeled by unconstraining the nodes along the root of the notch over 2 mm further down. The modes of the beam were computed using free-free boundary conditions. The transfer functions were computed using Eq. (8). The Q-factor values could be varied from one simulation to another but for a given simulation all modes were assigned the same Q. The force input was applied at one end of the beam, along its axis, and the output was taken on the other end (at the location of the PZT discs in the experimental setup—the discs were not actually modeled). A typical computational result is shown in Fig. 9 where all modes were assigned a Q of 1000.

As before, the ratio of the first sideband amplitude over the carrier amplitude is plotted against the carrier frequency varied from 50 to 210 kHz. It is immediately apparent that the modal density is now much greater than that of Fig. 8. Even with a purely axial excitation, bending and other nonlongitudinal modes will be excited due to the asymmetry introduced by the notch. Resonances become even denser above 100 kHz, which is around the frequency of the first cross-section resonance of the beam (i.e., the cut-off frequency of the first higher order mode). This plot provides a more realistic frequency dependence of the value of the ratio with the ultrasonic probing frequency. It also shows that the ratio peaks at −7 dB throughout the frequency range, therefore a higher modal density does not affect the optimal sensitivity. This optimal sensitivity at −7 dB contrasts with the −17 dB obtained at best with the receptance analysis, see Fig. 8. This discrepancy can easily be explained by the crudeness of the defect modeling in the receptance analysis. Given this fundamental difference in the models, one would not expect the two analyses to yield predictions matching accurately. Figure 9 also shows that for higher modal densities, it becomes extremely difficult to correlate peaks in the ratio curve with individual resonances in the open/close transfer functions. At high frequencies, the identification of resonances itself is problematic because of the strong coupling between closely spaced modes.

Similar FE computations were carried out for varying equivalent defect sizes (from 1% to 10% cross-section area reduction) and Q-factors ranging from 100 to 1000. The result is shown in Fig. 10.

The quantity plotted along the vertical axis of this plot is the averaged ratio over the whole frequency range (50–210 kHz). It is clear that the damping in the system has a significant effect on the sensitivity. The average ratio reaches about −25 dB for a 10% cracked area and $Q=1000$. The sensitivity drops to about −40 dB for the same Q and a 1% crack depth and further drops to −50 dB for a 1% crack depth and $Q=100$. Recalling that the experimental "noise floor" given by the damaged index of undamaged specimens was around −65 dB, these simulation results suggest that it should be theoretically possible to detect crack smaller than 1%. However in this simple model the strain dependence of the vibro-modulation effect cannot be explained and this theoretical sensitivity performance must be reviewed when the effect of residual stresses or initial crack opening are taken into account.

C. Case with compressive stresses

1. Introduction

In the case with no residual stresses studied so far, the crack faces are either completely mating or not at all in contact: effectively the crack tip acts like a hinge. In many real-life situations there is likely to be some residual stress in the cracked region.[16] The clear strain dependence observed in the experimental results (see Figs. 4 and 5) also indicates that the cracks do not behave as simple on/off switch. When compressive stresses are present around the crack, the tensile bending stresses will have to overcome the residual compressive ones in order for the crack to open. The variation of crack closure with time is no longer a simple square function since it then depends on the instantaneous value of the bending strain. The crack faces gradually mate over an increasingly larger area thus progressively reducing the defect size so that the breathing is smooth. Intuitively it is clear that the effect of this type of breathing on the modulation is that, depending on the relative value of the static stresses compared to the bending stresses, the effective crack size can be significantly smaller than the actual size.[16,17] This is confirmed by the following simplified analysis.

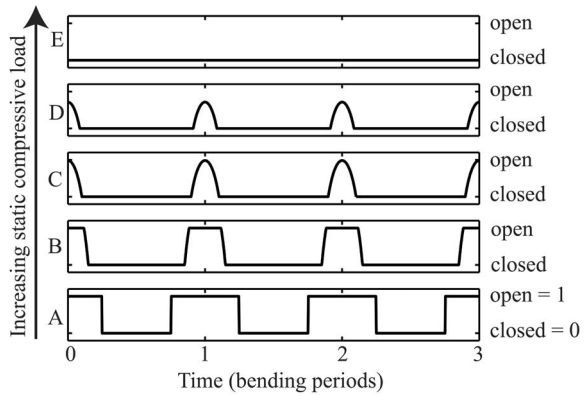


FIG. 11. Variation of opening of the crack with time for varying magnitudes of static compressive load. This load is very light in case A and is increased gradually from cases A to E. The opening at a given time is represented by a number between 0 (fully closed) and 1 (fully open). The time is given in multiples of the bending mode period T.

2. Modeling of crack closure in the presence of a static compressive stress

Consider a beam under a static compressive load P, also subjected to a sinusoidally oscillating bending moment. The resultant stress is then given by[26]

$$\sigma = \frac{yM_0 \sin(2\pi t/T)}{I} - P/A, \tag{9}$$

where A is the cross-section area of the beam, I its second moment of area, y the distance from the neutral axis, and M_0 the amplitude of the bending moment, which oscillates with a period T. M_0 was chosen so that it induces a strain of 100 $\mu\varepsilon$ at the top surface of a beam whose properties are the same as those assumed for the experimental specimens. Therefore the values of P obtained through this analysis are plausible experimental compressive stresses for the beams considered here. The crack was considered open when the resultant stress becomes tensile within the cracked region (here taken as the top 10% of the beam height). Figure 11 presents some results obtained from this model.

This set of plots shows the degree of opening and closing of the crack as the bending moment goes through three periods of oscillation T. The opening is represented by a number between 0 (closed) and 1 (fully open). Each subplot, labeled from A to E corresponds to a different value of the static compressive load. The particular cases shown were chosen because they illustrate the various possibilities of crack closure allowed by the model—the numerical values used to obtain them are of no particular interest at this stage. Plot A shows the result obtained when the static load is very low. The crack then actually behaves as described in the previous section for a cracked region free of residual stresses. At the other extreme, case E corresponds to a very large value of the compressive load. The static stress is then so large that it cannot be overcome by the tensile bending stress; consequently the crack remains closed. Plots B–D are intermediate cases. In case B, the static compressive stress is low enough to allow the crack to open fully but the time it stays open is shortened compared to case A. In case D, the crack can never open completely. Case C is the limiting case between B and D where the crack just reaches full opening but does not stay open for a finite duration.

3. Influence of compressive stresses on the previous results

To combine these results with the model presented so far requires a slightly more general formulation. Case D will be treated as an example because it possesses the interesting feature that the crack is never fully open. According to the model, the transfer function shift is directly related to the change in crack closure; therefore, one can expect the shape of the modulating function to follow that of case D scaled so that H_o is the value reached when the crack is at its maximum (but partial) opening and H_c is the value when the crack is closed.

Again defining X_m as the Fourier transform of the modulating function, X_m can be written

$$X_m = c_0 + c_1\delta_{f_m} + c_2\delta_{2f_m} + \ \dots\ , \tag{10}$$

where the c_k are the Fourier coefficients. Convolving X_m with a pure ultrasonic sine wave of amplitude A and frequency f_0 yields

$$X_0 * X_m = \underbrace{Ac_0\delta_{f_0}}_{\text{Carrier}} + \underbrace{Ac_1(\delta_{f_0+f_m} + \delta_{f_0-f_m})}_{\text{First sidebands}} + \ \dots\ . \tag{11}$$

The ratio, R, of the first sideband amplitude over the carrier amplitude is therefore

$$R = \frac{\text{First Sideband Amplitude}}{\text{Carrier Amplitude}} = \frac{c_1}{c_0}. \tag{12}$$

In other words this is the ratio of the first two Fourier coefficients of the modulating function. This is in agreement with the results obtained previously when the modulating function was a square wave.

To couple this formal result with an actual beam model, the receptance analysis was used again as the previous section showed that it captures all important features at a much lower computational cost. The beam was again split into three, with a short middle beam of reduced cross section representing the crack. The fully open crack was represented by cross-section area reduction of 10%. This models a uniform crack depth of 10% of beam depth. As before, the notch was also represented by an initial cross-section area reduction of 10% applied to all cases.

In case D in Fig. 11, the crack opens over 70% of its total depth. If the *total* crack depth is 10% of the beam depth, this means that the *effective* crack depth is 7%. Once this is known, it is possible to compute the transfer function for a beam with a 7% crack depth; this transfer function is then used for $H_o(f)$. As before, $H_c(f)$ is computed for the reference, notched but uncracked beam. Once these two transfer functions are known, the modulating function can be scaled appropriately for each ultrasonic frequency. The sideband/carrier ratio is then obtained from the coefficients of the numerical FFT of this scaled modulating function. This procedure can be repeated for various values of the static compressive load encompassing all the cases shown in Fig.

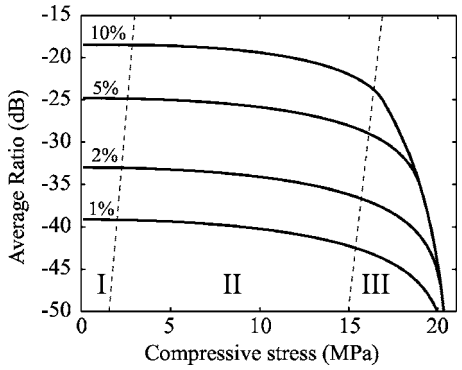


FIG. 12. Averaged ratio vs varying compressive stress. The various curves show the variation of the averaged ratio for a given total crack depth indicated above each curve as a percentage of beam cross-section area. This plot shows how the compressive stresses affect the crack opening and consequently the sideband amplitude. In region I, the beam effectively behaves as if there was no compressive stresses. In region II, the crack opens fully but the time it stays open decreases gradually as the compressive stress increases. Region III corresponds to cases where the crack only partially opens. All these curves were obtained for $Q=1000$.

11. The result from such an analysis is shown in Fig. 12, which shows the averaged vibro-modulation ratio against the static compressive load.

The different solid line curves show the variation of the averaged ratio for various total crack depths, indicated above each curve as a percentage of the beam cross-section area. All these curves were obtained for $Q=1000$. Two dashed lines define three regions, labeled I, II, and III in Fig. 12. In region I, the beam effectively behaves as if there was no compressive stresses: the compressive stress is not large enough to alter significantly the square modulating function. Therefore the average ratio remains constant at around -18 dB for the top curve which was the value previously obtained with no residual stresses at an equivalent crack depth of 10%, and $Q=1000$. Case A in Fig. 11 belongs to this region. In region II, the crack opens fully but the time it stays open decreases gradually as the compressive stress increases. In this region, the two transfer functions corresponding to the two extreme states are always the same: for the closed state it is still the reference $H_c(f)$ and in the open state it is $H_o(f)$ for a crack opened over its full depth. Within this range of stress, the slow decrease in the ratio is solely due to the shortening of the time the crack remains fully open. Therefore, although the two extreme transfer functions are the same, the average ratio decreases slowly from -18 to -25 dB with increasing compressive stress. This type of behavior is exemplified by case B. Cases A and B in Fig. 11 have different Fourier transforms although the crack opens fully in both cases. Region III corresponds to cases where the crack only partially opens as in case D. As can be expected, the averaged ratio then drops severely: as the stress increases in this region a point is reached at which the crack no longer opens. The average ratio then becomes minus infinity (this was cut off from Fig. 12 to allow other features to be seen more clearly).

This analysis provides an insight into the effect of the level of bending strain necessary to observe a vibro-modulation effect. Some residual compressive stresses are often present around fatigue cracks. The strain induced by the low frequency vibration will then have to be large enough to overcome these compressive stresses. This has important practical consequences since a defect will not be detected if the bending strain threshold is not reached. If the threshold is reached but the low frequency strain is not large enough to allow the crack to open fully, then the system effectively behaves as if the crack size had been reduced, thus resulting in a loss of sensitivity.

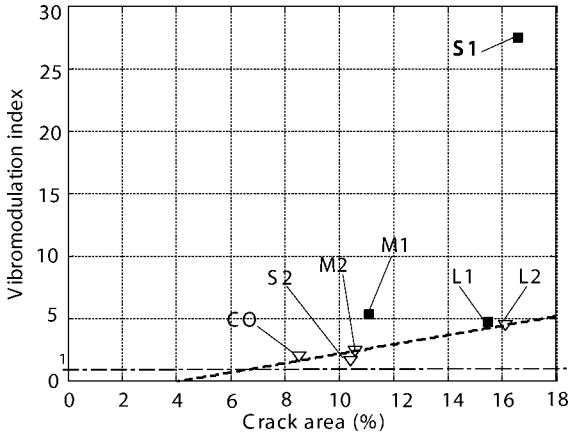


FIG. 13. Vibro-modulation damage index [see Eq. (1)] plotted against the crack area. The dashed line represents the reference value 1 for the undamaged specimens.

4. Outline of modeling of partially open cracks

Other crack features can affect the modulation process. An important case is when the crack is partially open while the beam is unloaded. Intuitively, it is clear that the beam will have to bend further to cause the crack to be fully closed.

One simple way to model this feature is to change the sign of the static stress P in the previous analysis. The residual stress then becomes tensile. The conclusions concerning the crack opening are similar, reversing the role of tensile and compressive stresses: it is necessary to reach a threshold of compressive bending stress for the crack to start closing and the modulation effect to be observed, thus resulting again in a loss of sensitivity.

IV. DISCUSSION

A. Comparison between damage index and crack size

Figure 13 shows the damage index plotted against the percentage of actual cross section covered by the crack (see Table I). The damage index appears to increase with the severity of the damage but there is considerable scatter in the normalized experimental data. For specimens L2, M2, S2, and CO, it is likely that the initial opening of the corresponding cracks is similar (as discussed earlier in Sec. II A). A trend-line was fitted to the experimental data for these specimens. This allows the smallest detectable crack using the IHS rig to be estimated by the intersection of this trend-line with the horizontal line at damage index=1, representing the noise floor of the testing rig. From this it appears that a cracked area of 7% is the smallest that can be expected to be detected by this implementation of the technique. The speci-

men S1 shows a much higher damage index compared to specimens L1 and L2 even though the crack is of very similar dimensions. This could be due to a different initial degree of closure of the crack that could result in an "effective" percentage crack size bigger than that of specimens L1 and L2. The meaning of such an "effective" crack area and its estimate is discussed next.

B. Comparison of experimental results with predictions

1. Comparing the experimental results with the model ignoring residual stresses

As much of the simulation work was carried out with a cracked area at 10% of the beam cross-section area, the experimental and theoretical results will first be compared for specimen M2 whose cracked area was measured to be 10.6%. The average ratio level for this specimen ranged from −63 dB at 10 $\mu\varepsilon$ to −55 dB at 50 $\mu\varepsilon$ (see Fig. 5). On the other hand, for an unstressed 10% crack size, the ratio obtained by FE modeling averaged over the frequency range 50–200 kHz ranges from −23 dB for $Q=1000$ to −34 dB for $Q=100$ (see Sec. III B 3). Similar levels of discrepancies obtain for all the beams tested so far. This further confirms the need to investigate more closely the state of opening of the crack when the beam is unloaded. For this purpose, another specimen, "CO" (for 'Crack Opening') was produced.

2. Estimate of the initial partial opening of the cracks

As mentioned earlier, the specimen "CO" was manufactured after the main series of tests on the other six beams had been completed, when it became apparent that the initial partial crack opening was likely to be significant.

A compliance method[27] was used in order to estimate the initial partial opening when the beam was unloaded. The principle exploited by this technique is that in the region close to the crack the stress/strain characteristic varies as the crack changes its openness under the effect of loading. This is due to local load path variation around the crack as the geometry (openness) of the crack changes.[28] Such path deviation would not occur once either of the two limit states of the crack are reached, when the crack is fully open or fully closed. The compliance test again used the three-point bending rig of Fig. 1. The crack opening behavior was measured by testing with the crack on the bottom face of the beam (this configuration is the one shown in Fig. 1), while closure was measured by reversing the beam so that the crack was on the top face and subject to compression when the beam was loaded. The strain characteristic in the proximity of the crack was measured by a strain gage (Kyowa model KFG 120) that was positioned across the crack itself (cross-crack strain gauge) whereas a reference strain gauge was positioned at the back of the beam with respect to the crack but on the same cross section (back-face strain gauge). The test could be conducted quasistatically or dynamically.[29] The loading of the specimen "CO" was conducted at 5 Hz and the load and strain signals for the case of loading applied to obtain crack closure or opening were monitored. Plotting the backface strain against the load produces the expected narrow ellipse, provided the load applied does not exceed the linear limit. This is shown in the upper curve of Fig. 14 which corresponds to loading with the crack on the upper face of the beam in Fig. 1, so that increased loading tended to close the crack. The hysteresis observed is essentially due to damping. If the cross-crack strain is plotted against the load a straight ellipse cannot be obtained since the crack closure or opening changes the stress/strain characteristic. The results are shown in the lower curve in Fig. 14. The comparison of the signals between the cross-crack and backface gauges demonstrates that only the cross-crack gauge is sensitive to the "breathing" of the crack because of the change of compliance (i.e., slope) in the strain versus load plot. In the region between approximately −0.8 and −1.4 kN the crack is in transition between the full and partial closure states. Therefore a criterion has to be chosen for estimating the load at which the crack can be considered closed. This closure load can be defined[29] by the intersection of the prolongation of the lines corresponding to the partial crack opening compliance, from 0 to −0.8 kN, and the full crack closure compliance below −1.4 kN (see the dashed lines in Fig. 14).

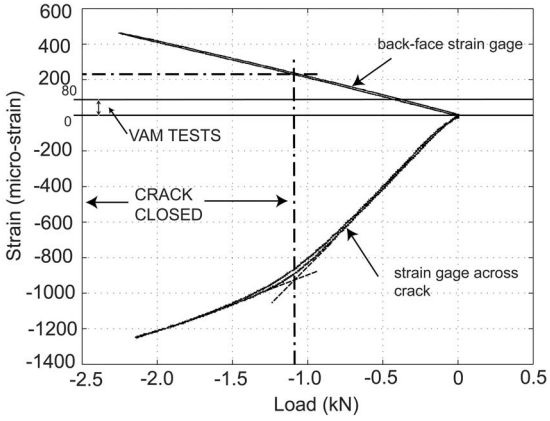


FIG. 14. Plot of strain against load for the backface and cross-crack strain gauges during loading of the specimen "CO" for crack closure. The full closure of the crack corresponds to an increased stiffness compared to partial closure. The exact closure point was estimated at −1.1 kN corresponding to 220 micro-strain measured at the backface strain gauge (dash-point line). The interval 0–80 microstrain at the backface strain gauge is highlighted since this was typically used during the VAM tests.

The compliance method is sensitive to the testing parameters,[29] however it proved to be adequate for estimating the degree of opening of the crack. Figure 14 shows that a compressive load of about −1.1 kN corresponding to a backface strain of about 220 $\mu\varepsilon$ was necessary to fully close the crack. The same experiment conducted with the beam turned over for opening the crack showed that the crack was fully open at a load of about −0.1 kN with a backface strain around −30 microstrain. Since the full opening load (−0.1 kN) is smaller than the full closure load (−1.1 kN), it could be concluded that the interfaces of the cracks produced in the specimens M1, L2, M2, S2, and CO were partially open at the end of fatiguing. The effect of this on the ultrasound modulation during the VAM tests is discussed next.

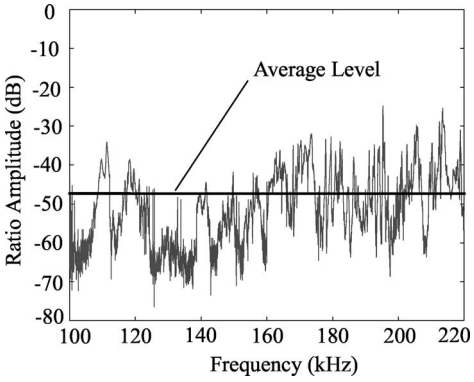


FIG. 15. Sideband-carrier ratio amplitude curve for the specimen "CO" at 80 microstrain average low-frequency vibration level as recorded by the backface strain gauge.

3. Comparing the experimental results with the model including the effect of residual stresses

In the previous section, it was shown that for the specimen CO, a strain of 220 $\mu\varepsilon$ at the backface was required to close the crack fully, and -30 $\mu\varepsilon$ at the backface resulted in the crack opening fully. The maximum low frequency strain just after impact was on average about ± 80 $\mu\varepsilon$ for specimen "CO." Therefore the degree of crack opening was changing between -30 and $+80$ $\mu\varepsilon$. For the crack state to vary between fully open and fully closed, the strain would need to vary between (at least) -30 and $+220$ microstrains. Therefore the fraction of the crack area contributing to the modulation was $(30+80)/(30+220)=0.44$. Since the measured crack area was 8.5% for "CO," the effective crack area was about 3.7%. For a crack of this size the finite element model predicted ratio levels around -45 dB, as can be seen in Fig. 10, for a Q of a few hundreds. The Q-factors of the isolated peaks around 100 kHz were measured for the CO specimen. The values were mostly below 100 with a few reaching 200 to a maximum of 500. The experimental sideband-carrier ratio for this specimen is shown in Fig. 15. The average ratio for an average strain of 80 $\mu\varepsilon$ was measured at -48 dB. This value is similar to the predicted value of -45 dB, provided the *effective* crack size is considered. Given the simplicity of the models, the level of agreement is encouraging.

V. CONCLUSIONS

In this paper it has been shown that the engineering of the testing rig is crucial for the reliable detection of cracks reliably using the VAM method. Appropriate supports and excitation devices have to be used in order to allow the modulation of the probing ultrasound (i.e., the sidebands in the frequency domain) due to the crack to be detectable. The supports and/or any other contact between the specimen and other external bodies are very likely to introduce sidebands at the same levels as those generated by a crack, and this could jeopardize the successful application of the VAM NDT method. It was found that the impact excitation setup gave sidebands about -65 dB below the carrier amplitude for an undamaged specimen and this was satisfactory for crack detection.

A damage index was defined and its value calculated for each of the laboratory beams using the impact-based rig. While the value of the damage index generally increased with crack size, the relationship was far from linear. This is likely to be due to different crack morphologies or initial degree of crack opening/closure due to the different fatigue conditions used. The trend obtained suggests that the smallest cracks that could be detected must be at least 7% of the cross-section area. Therefore the technique will not be able to detect small cracks reliably.

A simple model was proposed and the overall level of the sidebands predicted was in broad agreement with the sideband level measured experimentally, provided the state of static opening of the crack is considered carefully. This model explains the occurrence of sidebands in the ultrasound signal by the geometrical changes undergone by the cracked region as the specimen is subjected to the low frequency vibration. It also accounts for the dependence of VAM effect on the ultrasonic frequency and the low frequency strain dependence when the crack is initially open or closed (due to geometry or to tensile/compressive stresses). Finally, it gives some insight into the importance of damping for this technique and provides approximate predictions of sideband levels for a given crack size and Q-factor.

Given the sensitivity to cracks obtained and the time required to sweep the ultrasonic frequency in order to ensure that a peak of sensitivity is obtained, this implementation of the vibro-acoustic modulation technique does not appear to offer significant potential for rapid screening for the detection of small cracks. However, other vibro-modulation implementations (e.g., using SAW) may be more sensitive.

[1]Y. Zheng, R. G. Maev, and I. Y. Solodov, "Nonlinear acoustic applications for material characterization: A review," Can. J. Phys. **77**, 927–967 (1999).
[2]P. B. Nagy, "Fatigue damage assessment by nonlinear ultrasonic material characterization," Ultrasonics **36**, 365–381 (1998).
[3]I. Y. Solodov, "Ultrasonics of nonlinear contacts: Propagation, reflection and nde-applications," Ultrasonics **36**, 383–390 (1998).
[4]N. Krohn, R. Stoessel, and G. Busse, "Acoustic nonlinearity for defect selective imaging," Ultrasonics **40**, 633–637 (2002).
[5]I. Y. Solodov, N. Krohn, and G. Busse, "CAN: an example of nonclassical acoustic nonlinearity," Ultrasonics **40**, 621–625 (2002).
[6]J. G. Sessler and V. Weiss, "Crack detection apparatus and method," US Patent 3867836.
[7]A. S. Korotkov and A. M. Sutin, "Modulation of ultrasound by vibration in metal constructions with cracks," Acoust. Lett. **18**, 59–62 (1994).
[8]A. S. Korotkov, M. M. Slavinskii, and A. M. Sutin, "Variation of acoustic nonlinear parameters with the concentration of defects in steel," Acoust. Phys. **40**, 84–87 (1994).
[9]V. Y. Zaitsev, A. M. Sutin, I. Y. Belyaeva, and V. E. Nazarov, "Nonlinear interaction of acoustical waves due to cracks and its possible usage for cracks detection," J. Vib. Control **1**, 335–344 (1995).
[10]D. M. Donskoy and A. M. Sutin, "Vibro-acoustic modulation nondestructive evaluation technique," J. Intell. Mater. Syst. Struct. **9**, 765–771 (1998).
[11]A. E. Ekimov, I. N. Didenkulov, and V. V. Kazakov, "Modulation of torsional waves in a rod with a crack," J. Acoust. Soc. Am. **106**, 1289–1292 (1999).
[12]V. Y. Zaitsev and P. Sas, "Nonlinear response of a weakly damaged metal sample: A dissipative modulation mechanism of vibro-acoustic interaction," J. Vib. Control **6**, 803–822 (2000).
[13]K. E.-A. V. D. Abeele, P. A. Johnson, and A. M. Sutin, "Nonlinear elastic wave spectroscopy (news) techniques to discern material damage. i. Nonlinear wave modulation spectroscopy (nwms)," Res. Nondestruct. Eval. **12**, 17–30 (2000).

[14]D. M. Donskoy, A. M. Sutin, and A. Ekimov, "Nonlinear acoustic interaction on contact interfaces and its use for nondestructive testing," NDT Int. **34**, 231–238 (2001).
[15]I. Y. Balyaeva and V. Y. Zaitsev, "Nonlinear scattering of acoustic waves by discontinuity-like defects in application to crack detection," J. Vib. Control **2**, 465–478 (1996).
[16]J.-Y. Kim, V. A. Yakovlev, and S. I. Rokhlin, "Parametric modulation mechanism of surface acoustic wave on a partially closed crack," Appl. Phys. Lett. **82**, 3203–3205 (2003).
[17]J.-Y. Kim, V. A. Yakovlev, and S. I. Rokhlin, "Surface acoustic wave modulation on a partially closed fatigue crack," J. Acoust. Soc. Am. **115**, 1961–1972 (2004).
[18]R. D. Adams and P. Cawley, in *Research Techniques in Nondestructive Testing*, edited by R. S. Sharpe (Academic, London, 1985), Vol. 8, Chap. 7, pp. 304–360.
[19]K. L. Narayan and C. Jebaraj, "Sensitivity analysis of local/global modal parameters for identification of a crack in a beam," J. Sound Vib. **228**, 997–994 (1999).
[20]G. M. Owolabi, A. S. J. Swamidas, and R. Seshadri, "Crack detection in beams using changes in frequencies and amplitudes of frequency response functions," J. Sound Vib. **265**, 1–22 (2003).
[21]B. W. Drinkwater, R. Dwyer-Joyce, and P. Cawley, "A study of the interaction between ultrasound and a partially contacting solid-solid interface," Proc. R. Soc. London, Ser. A **452**, 2613–2628 (1996).
[22]J. M. Richardson, "Harmonic generation at an unbounded interface. i. planar interface between semi-infinite elastic media," Int. J. Eng. Sci. **17**, 73–85 (1979).
[23]R. E. Ziemer and W. H. Tranter, *Principles of Communication* (Wiley, New York, 2003).
[24]R. E. D. Bishop and D. C. Johnson, *The Mechanics of Vibration* (Cambridge University Press, Cambridge, 1960).
[25]L. Meirovitch, *Elements of Vibration Analysis* (MacGraw-Hill, New York, 1986).
[26]S. H. Crandall, N. C. Dahl, and T. J. Lardner, *An Introduction to the Mechanics of Solids* (McGraw-Hill, Tokyo, Japan, 1978).
[27]S. Stoychev and D. Kujawski, "Methods for crack opening load and crack tip shielding determination: A review," Fatigue Fract. Eng. Mater. Struct. **26**, 1053–1067 (2003).
[28]Y. Verreman, "Monitoring short fatigue cracks with miniature strain gages," Exp. Mech. **34**, 207–216 (1994).
[29]V. Sarma, G. Jaeger, and A. Koethe, "On the comparison of crack closure evaluation using dynamic and static compliance measurements," Int. J. Fatigue **23**, 741–745 (2001).

Experimental and numerical investigations on resonant characteristics of a single-layer piezoceramic plate and a cross-ply piezolaminated composite plate

Chien-Ching Ma[a)]
Department of Mechanical Engineering, National Taiwan University, Taipei, Taiwan 106, Republic of China

Hsien-Yang Lin[b)]
Department of Mechanical Engineering, De Lin Institute of Technology, Tu-Cheng, Taiwan 236, Republic of China

Yu-Chih Lin
Department of Biomedical Engineering, Yuanpei Institute of Science and Technology Hsinchu, Taiwan 300, Republic of China

Yu-Hsi Huang
Department of Mechanical Engineering, National Taiwan University, Taipei, Taiwan 106, Republic of China

(Received 10 June 2004; revised 19 October 2005; accepted 23 November 2005)

Piezolaminated composite plates are widely used in many industrial applications such as intelligent structures and advanced aerospace structural applications. To improve the dynamic performance of the piezolaminated composite structures, it is necessary to experimentally investigate the resonant characteristics of these structures. Three measurement techniques are used in this study to investigate the vibration behavior of tested specimens. The first method, AF-ESPI (amplitude-fluctuation electronic speckle pattern interferometry), is the major technique for measuring the resonant characteristics of a single-layer piezoceramic plate and a multilayer cross-ply glass fiber reinforced plastics piezolaminated composite plate for completely free conditions. The completely free condition is approximated in experiment by placing and partially sticking the specimen on the surface of a very soft sponge. A thin single-layer piezoceramic plate is first investigated and up to twenty-six transverse modes are presented. Excellent quality of interferometric fringe patterns for vibration mode shapes is presented. The influence of the actual boundaries on the resonant characteristics of the thin piezoceramic plate is discussed. The second method, laser Doppler vibrometer, and the third method, impedance analyzer, are both employed to verify the AF-ESPI results for the piezolaminated composite plate. Both in-plane and out-of-plane resonant frequencies and vibration mode shapes of the piezolaminated composite plate are demonstrated. Finally, numerical computations based on the finite element analysis are presented for comparison with the experimental results. Excellent agreement between the experimental measured data and the numerical calculated results are found for resonant frequencies and mode shapes of the single-layer piezoceramic plate and the cross-ply piezolaminated composite plate. [DOI: 10.1121/1.2161428]

PACS number(s): 43.40.At, 43.40.Dx [CBB] Pages: 1476–1486

I. INTRODUCTION

Piezoelectric materials create electric charge when mechanically stressed, conversely, a change in dimensions will be induced upon application of the electric field. Thus, the piezoelectric effect that expresses the connection of electrical and mechanical fields has the built-in performance of the sensing and actuating capacities simultaneously. Over the past few decades, there has been considerable interest in the use and applications of piezoelectric materials in electromechanical transducers, underwater acoustic, electronic packaging, resonators, frequency-band filter, vibration damping, and smart structures. One class of piezoelectric materials is the lead zirconate-titanate (PZT) based ceramics. The use of PZT in piezolaminated composite structures has received a great deal of attention because the sensing and actuating capacities become part of the structures in piezolaminated composites, and the ease of integrating piezoceramic materials by means of embedding those to laminated structures. These advantages enhance the growing interest in fast applications of piezolaminated composite structures in the areas of vibration actuation, control, and load/health monitoring. Generally, the piezolaminated composite structures are fabricated by incorporated PZT or polyvinylidene difluoride (PVDF) into piezolaminated composites with fiber-reinforced layers. By combining many distinct layers, the

[a)]Author to whom correspondence should be addressed. Electronic mail: ccma@ntu.edu.tw
[b)]Electronic mail: sylin@dlit.edu.tw

advantages of each ply can be developed more effectively. Hence, the study of embedded piezoelectric materials in composites has received considerable attention in recent years. However, several researchers have carried out the development of mathematical models and associated computational models to predict the response of piezolaminated composites. Ha *et al.*[1] used an eight-node composite brick element and variational principle to construct the FEM formulation for modeling the dynamic and static response of laminated composites containing distributed piezoceramics subjected to both mechanical and electrical loading. Batra and Liang[2] used the three-dimensional linear theory of elasticity to study the vibrations of simply supported rectangular laminated plate with embedded PZT layers, numerical results for a thin and a thick plate with one embedded actuator layer and one embedded sensor are presented. Abramovich[3] presented an exact elasticity solution for the forced induced vibrations of a piezolaminated elastic beam. This analysis was carried out by using the method of Fourier series and the solution was exact within the assumptions of linear elasticity and plane strain deformation. Loja *et al.*[4] used a family of higher order B-spline finite strip models for the static and free vibration analysis of laminated plates with arbitrary shapes and lay-ups, loading and boundary conditions. The above-mentioned references were usually presented by the theoretical analysis or the numerical computation. Few results used experimental technique to study the vibration characteristics of piezolaminated composite plates. Moreover, the results obtained by experimental methods are always confined in the first few modes. The lack of experimental investigation combined with numerical analysis of resonant characteristics for single-layer piezoceramic plate and the piezolaminated composite plate with an embedded PZT layer in cross-ply fiber reinforced composites are the direct motivation for the study presented herein. For the case of the piezolaminated composite plate, the resonant characteristics of in-plane and out-of-plane motions are investigated. Two noncontact and real-time laser interferometer optical methods, the full-field electronic speckle pattern interferometry (ESPI) technique and the pointwise laser Doppler vibrometer (LDV) technique, are used in this study.

ESPI, which utilizes the information of speckle phenomenon on an optically rough surface, is a full-field and noncontact laser-based optical metrology technique for real-time measurement of surface displacement. Because ESPI excels in the flexibility, real-time, and full-field measurement, it becomes an exciting and highly active field in optical metrology. Furthermore, the image data are digitized by a video camera and processed by a digital signal processing system for the ESPI method, which eliminates the slow and cumbersome process of film development used for conventional technique. The comparative advantage of operation allows ESPI to extend its applications to many practical engineering fields compared with other optical measurement techniques. The application of ESPI to the analysis of vibration study can be traced back to the work by Butters and Leendertz[5] during the 1970s. The detailed description of ESPI techniques can be found in the book written by Jones.[6] Now, ESPI is a well-established technique for vibration analysis due to ESPI using a video recording and display in real time. The most convenient and widely used experimental setup of vibration measurement by ESPI is the time-averaging method that yields a video image of vibrating object with correlation fringes superimposed. The fringe patterns can be expressed by the zero-order Bessel function and correspond to the contour of constant vibration amplitude or surface displacement. However, the main restrictions of the time-averaging method are poor image quality, decreased visibility with vibration amplitude, and the lack of the information on vibration phase. Numerous research further developing on the ESPI method has been addressed to overcome these limitations and to extend its usage in practical applications. Hogmoen and Løkberg[7] proposed a method that was based on sinusoidal phase modulation of reference wave in time-averaging ESPI for providing information of the vibration phase. Nakadate[8] used four phase-shifted speckle interferograms for extraction of numerical data of vibration amplitude. Valera *et al.*[9] described a stroboscopically illuminated ESPI system by modulating both the illuminating pulse phase relative to the vibration and the optical phase to the interferometer reference beam to generate eight video frames which were processed to generate the vibration phase. Wong[10] improved the fringe contrast and sensitivity by subtracting two Bessel fringe patterns at two different force levels. Wang *et al.*[11] proposed the amplitude-fluctuation ESPI (AF-ESPI) technique that was based on video-signal-subtraction, but the reference image was taken from a vibrating state instead of a free state. The fringe patterns produced by AF-ESPI method have enhanced visibility and reduced noise. Wang *et al.*[12] further employed the modal testing technique and the AF-ESPI technique to measure the mode shapes and resonant frequencies of the shadow mask in a multimedia monitor. Ma and co-worker[13] further developed the AF-ESPI based on the work of Huang and Ma[14] and impedance analyzer for vibrations analysis of piezoelectric materials.

The aim of this work is to investigate experimentally the resonant characteristics of a thin single-layer piezoceramic plate and a piezolaminated composite plate with an embedded piezoceramic layer in cross-ply glass fiber reinforced composites for the completely free boundary condition. The experimental techniques used in this study including the AF-ESPI, LDV, and impedance analyzer. From the AF-ESPI method, the resonant frequencies and mode shapes of a single-layer piezoceramic plate and a piezolaminated composite plate with free boundary condition are obtained. Up to 26 transverse modes for the thin single-layer piezoceramic plate are presented. The quantitative magnitudes of the full-field displacements are also indicated in the experimental results for each mode, and the sensitivity is submicrometer. The resonant mode shapes of in-plane and out-of-plane vibrations of the piezolaminated composite plate are also demonstrated. To verify the AF-ESPI results, the LDV, which is a pointwise displacement measurement technique, is also used as the second experimental technique. The third experimental technique, the impedance analyzer, is employed to determine resonant frequencies of in-plane vibration for the piezolami-

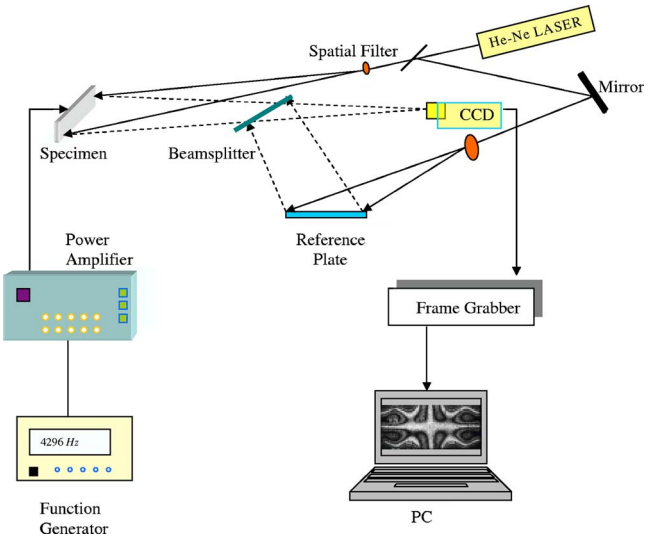


FIG. 1. (Color online) Schematic diagram of AF-ESPI for out-of-plane measurement.

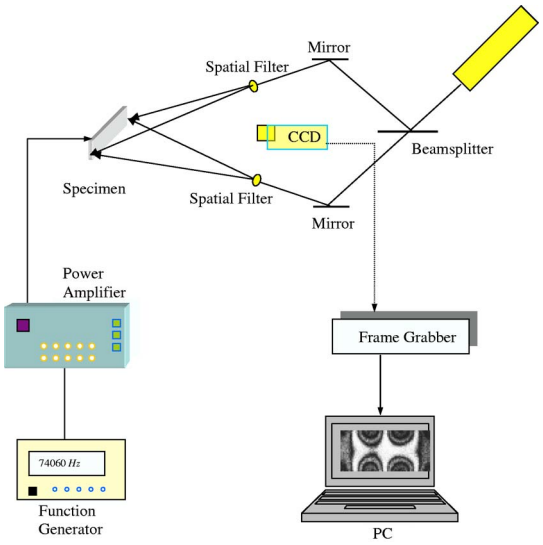


FIG. 2. (Color online) Schematic diagram of AF-ESPI for in-plane measurement.

nated composite plate. Furthermore, in light of the results presented in this work, numerical computation based on the finite element method (FEM) utilizing ABAQUS commercial software package[15] are also made. In practical testing condition, the free boundary in experiment is approximated by placing and partially sticking the specimen on the surface of a very soft sponge. A detailed calculation is made to evaluate the effect of sponge on the vibration behavior of the thin single-layer piezoceramic plate. The excellent agreement of the numerical results in resonant frequencies and mode shapes with experimental measurements is found for the single-layer piezoceramic plate and the cross-ply piezolaminated composite plate.

II. THEORY OF AF-ESPI METHOD FOR OUT-OF-PLANE AND IN-PLANE MEASUREMENTS

The time-averaging method is the most commonly used technique of ESPI for experimental vibration measurement. The name "time-averaging" denotes that the vibration measurement includes many periods of object motions during the camera frame period. Two different optical setups (out-of-plane sensitivity and in-plane sensitivity) are employed in this study for the vibration measurement. The optical setups for out-of-plane and in-plane vibration measurements by AF-ESPI are shown schematically in Figs. 1 and 2, respectively. Fundamentals with regards to the AF-ESPI method for in-plane and out-of-plane measurements can be found in Refs. 13 and 14.

A. Out-of-plane vibration

When the specimen vibrates periodically, the light intensity of the vibrating image detected by a charge-coupled device (CCD) camera can be expressed by the time-averaging method as

$$I = \frac{1}{\tau}\int_0^\tau \left\{ I_O + I_R + 2\sqrt{I_O I_R}\cos\left[\phi + \frac{2\pi}{\lambda}A(1 + \cos\theta)\cos\omega t\right]\right\} dt, \quad (1)$$

where I_O is the object light intensity, I_R is the reference light intensity, ϕ is the random phase difference between the object field and the reference field, λ is the wavelength of laser, A is the vibration amplitude, θ is the angle between the object light and the observation direction, τ is the CCD refresh time, and ω is the angular frequency.

Assume that the vibration period is much shorter than the CCD refresh time (i.e., $\tau \gg 2\pi/\omega$) and let

$$\Gamma = \frac{2\pi}{\lambda}(1 + \cos\theta), \quad (2)$$

then, Eq. (1) can be approximated to a simplified form as indicated by Jones and Wyke[6]

$$I_1 = I_O + I_R + 2\sqrt{I_O I_R}(\cos\phi)J_0(\Gamma A), \quad (3)$$

where J_0 is a zero-order Bessel function of the first kind. As the vibration of specimen goes on, we take the second image by CCD and assume that the vibration amplitude of the second image has changed from A to $A+\Delta A$ due to the electronic noise or the instability of the apparatus. The light intensity of the second image can be represented as

$$I_2 = \frac{1}{\tau}\int_0^\tau \{I_O + I_R + 2\sqrt{I_O I_R}\cos[\phi + \Gamma(A + \Delta A)\cos\omega t]\} dt. \quad (4)$$

Expanding Eq. (4) by using the Taylor series expansion, keeping the first two terms and neglecting the higher order terms, we get

$$I_2 = I_O + I_R + 2\sqrt{I_O I_R}(\cos\phi)\left[1 - \tfrac{1}{4}\Gamma^2(\Delta A)^2\right]J_0(\Gamma A). \quad (5)$$

When these two images, i.e., I_1 and I_2, are subtracted and rectified by the image-processing system, the resulting light intensity can be expressed as

$$I = |I_2 - I_1| = \frac{\sqrt{I_R I_O}}{2} |(\cos\phi)\Gamma^2(\Delta A)^2 J_0(\Gamma A)|. \quad (6)$$

B. In-plane vibration

Similar to the out-of-plane vibration case, the first and second image intensities, i.e., I_1 and I_2, for in-plane vibration using the time-averaging method are expressed as

$$I_1 = I_O + I_R + 2\sqrt{I_O I_R}(\cos\phi) J_0(\Gamma' A'), \quad (7)$$

and

$$I_2 = I_O + I_R + 2\sqrt{I_O I_R}(\cos\phi)\left[1 - \tfrac{1}{4}\Gamma'^2(\Delta A')^2\right] J_0(\Gamma' A'), \quad (8)$$

where I_O is the object light intensity, I_R is the reference light intensity, A' is the vibration amplitude of in-plane vibration, $\Gamma' = 4\pi \sin\theta'/\lambda$, and θ' is half of the angle between two illumination lights. Subtracting Eq. (8) from Eq. (7) and rectifying by the image processing system, we can obtain the image intensity as

$$I = |I_2 - I_1| = \frac{\sqrt{I_R I_O}}{2} |(\cos\phi)\Gamma'^2(\Delta A')^2 J_0(\Gamma' A')|. \quad (9)$$

From Eqs. (6) and (9), it is interesting to note that the fringe patterns for both the out-of-plane and in-plane vibration motions obtained by AF-ESPI are modulated by a zero-order Bessel function J_0. Combining the out-of-plane and the in-plane optical setups by the AF-ESPI method, we can construct completely the vibration characteristics of piezoelectric materials, including resonant frequencies and mode shapes at the same time.

III. EXPERIMENTAL AND NUMERICAL RESULTS

Because the polarized piezoelectric ceramics have the same symmetry as a hexagonal crystal in class C_{6v} mm, it can be modeled as a transversely isotropic material. For a transversely isotropic material referred to a Cartesian coordinate system x_1, x_2, x_3 and assuming that $x_1 - x_2$ is the isotropic plane and x_3 is the poling direction, then the constitutive equation for piezoelectric ceramics can be represented in a matrix form as

$$\begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \\ T_6 \\ D_1 \\ D_2 \\ D_3 \end{bmatrix} = \begin{bmatrix} C_{11}^E & C_{12}^E & C_{13}^E & 0 & 0 & 0 & 0 & 0 & -e_{31} \\ C_{12}^E & C_{11}^E & C_{13}^E & 0 & 0 & 0 & 0 & 0 & -e_{31} \\ C_{13}^E & C_{13}^E & C_{33}^E & 0 & 0 & 0 & 0 & 0 & -e_{33} \\ 0 & 0 & 0 & C_{44}^E & 0 & 0 & 0 & -e_{15} & 0 \\ 0 & 0 & 0 & 0 & C_{44}^E & 0 & -e_{15} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & (C_{11}^E - C_{12}^E)/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & e_{15} & 0 & \varepsilon_{11}^s & 0 & 0 \\ 0 & 0 & 0 & e_{15} & 0 & 0 & 0 & \varepsilon_{11}^s & 0 \\ e_{31} & e_{31} & e_{33} & 0 & 0 & 0 & 0 & 0 & \varepsilon_{33}^s \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \\ E_1 \\ E_2 \\ E_3 \end{bmatrix}. \quad (10)$$

From Eq. (10), it is obvious that the complete description of the material characteristic for lead zirconate titanate (PZT) ceramics requires ten material constants; they are five elastic constants (C_{11}^E, C_{12}^E, C_{13}^E, C_{33}^E, and C_{44}^E), three piezoelectric constants (e_{15}, e_{31}, and e_{15}), and two dielectric constants (ε_{11}^s and ε_{33}^s). In the following, the resonant characteristics of a single-layer piezoceramic plate and a multilayer cross-ply piezolaminated composite plate are investigated and discussed in detail.

A. The results of the single-layer piezoceramic plate

The thin single-layer piezoceramic plate with completely free boundary is analyzed first. The plate has the geometrical dimension of $70 \times 25 \times 0.26$ mm and is made of $Pb(Zr \cdot Ti)O_3$ ceramics, the model number is PIC-151 (Physik Instrumente). The material properties of the thin piezoceramic plate are listed in Table I which are provided by Physik Instrumente. The polarization is parallel to the x_3 axis, and two opposite faces of the piezoceramic plate are completely coated with silver electrodes. The piezoceramic plate is excited by the application of an ac voltage across electrodes on the two surfaces and has the completely stress-free boundary conditions. Figure 3 is a schematic diagram of the specimen configuration. A self-arranged time-averaging ESPI system is used to perform the experimental measurements for resonant frequencies and corresponding full-field mode shapes. Since the piezoceramic plate is very thin, it will display large transverse (out-of-plane) motion compared to any possible in-plane components, especially in the low resonant frequencies. Hence only the optical setup for the out-of-plane sensitivity is used for the single-layer piezoceramic plate.

As shown in Fig. 1, a He–Ne Laser (Uniphase 1135P) with 30 mW and wavelength $\lambda = 632.8$ nm is used as the coherent light source. The emitting laser beam is split into two parts by a variable beamsplitter. One beam is directed toward the tested piezoceramic plate and then reflects to the CCD camera acting as the object beam. The second one which

TABLE I. Material constants of PIC-151.

Quality	PIC-151
C_{11}^E (N/m^2)	10.76×10^{10}
C_{12}^E	6.313×10^{10}
C_{13}^E	6.386×10^{10}
C_{33}^E	10.04×10^{10}
C_{44}^E	1.962×10^{10}
C_{66}^E	2.224×10^{10}
e_{31} (N/Vm)	−9.52
e_{33}	15.14
e_{15}	11.97
ρ (kg/m^3)	7800
$\epsilon_{11}^s/\epsilon_0$	1111
$\epsilon_{33}^s/\epsilon_0$	925
ϵ_0	8.85×10^{-12}

serves as a reference beam is illuminated on the surface of a reference plate and reflects into the CCD camera via the beamsplitter. The object and reference beams are combined into the CCD sensor array through a zoom lens (Nikon Micro-Nikkor 55 mm). It is important to note that the optical path and the light intensity of these two beams should remain identical in the experimental setup. A CCD camera (Pulnix TM-7CN) and a frame grabber (Dipix P360F) with a digital signal processor on board are used to record and process the images obtained from interferogram of the object and reference beams. Once the specimen vibrates, the interferogram recorded by the CCD camera is stored in an image buffer as a reference image. Then the next frame is grabbed and is subtracted by the image processing system. The CCD camera converts the intensity distribution of the interference pattern of the object into a corresponding video signal at 30 frames/s. The signal is electronically processed and finally converted into an image on the video monitor. The interpretation of the fringe image is similar to the reading of a displacement contour. To achieve a sinusoidal output, a digitally controlled function generator (Hewlett Packard, HP-33120A) connected to a power amplifier (NF Electronic Instruments 4005 type) is employed as an input source, which generates periodical exciting force to the specimen.

The detailed experimental procedure of the AF-ESPI technique is performed as follows. First, a reference image is captured after the specimen vibrates, then the second image is taken sequentially, and the reference image is subtracted by the image processing system. If the vibrating frequency is not the resonant frequency, only randomly distributed speckles are displayed and no fringe patterns will be shown. However, if the vibrating frequency is in the neighborhood of the resonant frequency, stationary distinct fringe patterns will be observed in the monitor. Then, the function generator is cautiously and gradually turned; the number of fringes will increase, and the fringe pattern will become clearer but the position of nodal lines will not change as the resonant frequency is approached. From the aforementioned experimental procedure, the resonant frequencies and the corresponding full-field mode shapes can be determined at the same time by using the AF-ESPI optical system. Because the thickness of the piezoceramic plate is very thin, it is easy to excite the

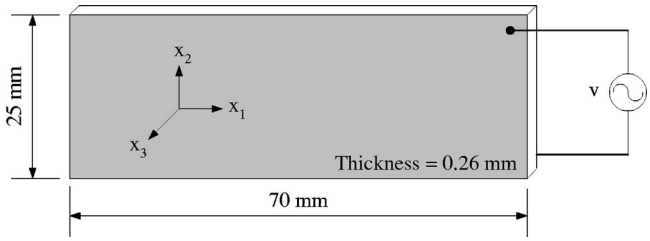


FIG. 3. Schematic diagram for the thin piezoceramic plate.

transverse vibration and up to 26 modes are obtained. The resonant frequencies are ranged from 221 to 5077 Hz.

In addition to the experimental measurement, numerical computations of resonant frequencies as well as mode shapes are also performed by a commercially available software ABAQUS finite element package; a 20-node three-dimensional piezoelectric element (C3D20RE) with reduced integration is used to analyze the problem. Generally, reduced integration provides more accurate results and significantly reduces computational time. Analytically, the boundary condition can be specified as completely free or as any constrained situation. The completely free boundary condition means that the specimen is, in fact, floating in space with no attachment or connection in ground and exhibits rigid body behavior at zero frequency. In testing practice, however, it is almost not realizable and generally not possible to fully achieve this condition. Hence, the specimen must be supported in some manner to approximate the completely free boundary. Since the optical setup as shown in Figs. 1 and 2 is put on a vibration isolated table, the perpendicular direction of the surface of specimen should be in the same plane of the table. It is not possible to hang the specimen in the air by using soft springs. In this study, the completely free boundary is approximated by sticking the single-layer piezoceramic plate onto the surface of a very soft sponge. Since the thickness of the piezoceramic plate is thin, the influence of sponge support on the vibration behavior for the out-of-plane modes is anticipated and should be discussed in detail. Hence, two different boundary conditions are employed in numerical calculations to study the influence of sponge support on the resonant frequencies and mode shapes. The first one is the completely free plate without any constraints or supports except that the bottom and top surfaces of the piezoceramic plate are set to be short-circuited. The second case includes the effect of an elastomeric sponge to model the actual experimental situation. The elastic behavior of highly compressible materials such as sponges or foams is based on the modified form of the strain energy function U,

$$U=\sum_{i=1}^{N}\frac{2u_i}{\alpha_i^2}\left[\hat{\lambda}_1^{\alpha_i}+\hat{\lambda}_2^{\alpha_i}+\hat{\lambda}_3^{\alpha_i}-3+\frac{1}{\beta_i}((J^{\mathrm{el}})^{-\alpha_i\beta_i}-1)\right], \tag{11}$$

where N is the order of strain energy potential, u_i, α_i, and β_i are material parameters, β_i is related to the Poisson's ratio by $\beta_i=\nu_i/(1-2\nu_i)$, $\hat{\lambda}_i=(J^{\mathrm{th}})^{-1/3}\ \lambda_i$ and λ_i are the principal stretches. J^{th} and J^{el} represent thermo and elastic volume ratio, respectively, and are defined as $J^{\mathrm{th}}=(1+\varepsilon^{\mathrm{th}})^3$, $J^{\mathrm{el}}=\lambda_1\lambda_2\lambda_3/J^{\mathrm{th}}$; $\varepsilon^{\mathrm{th}}$ is the linear thermal expansion strain that

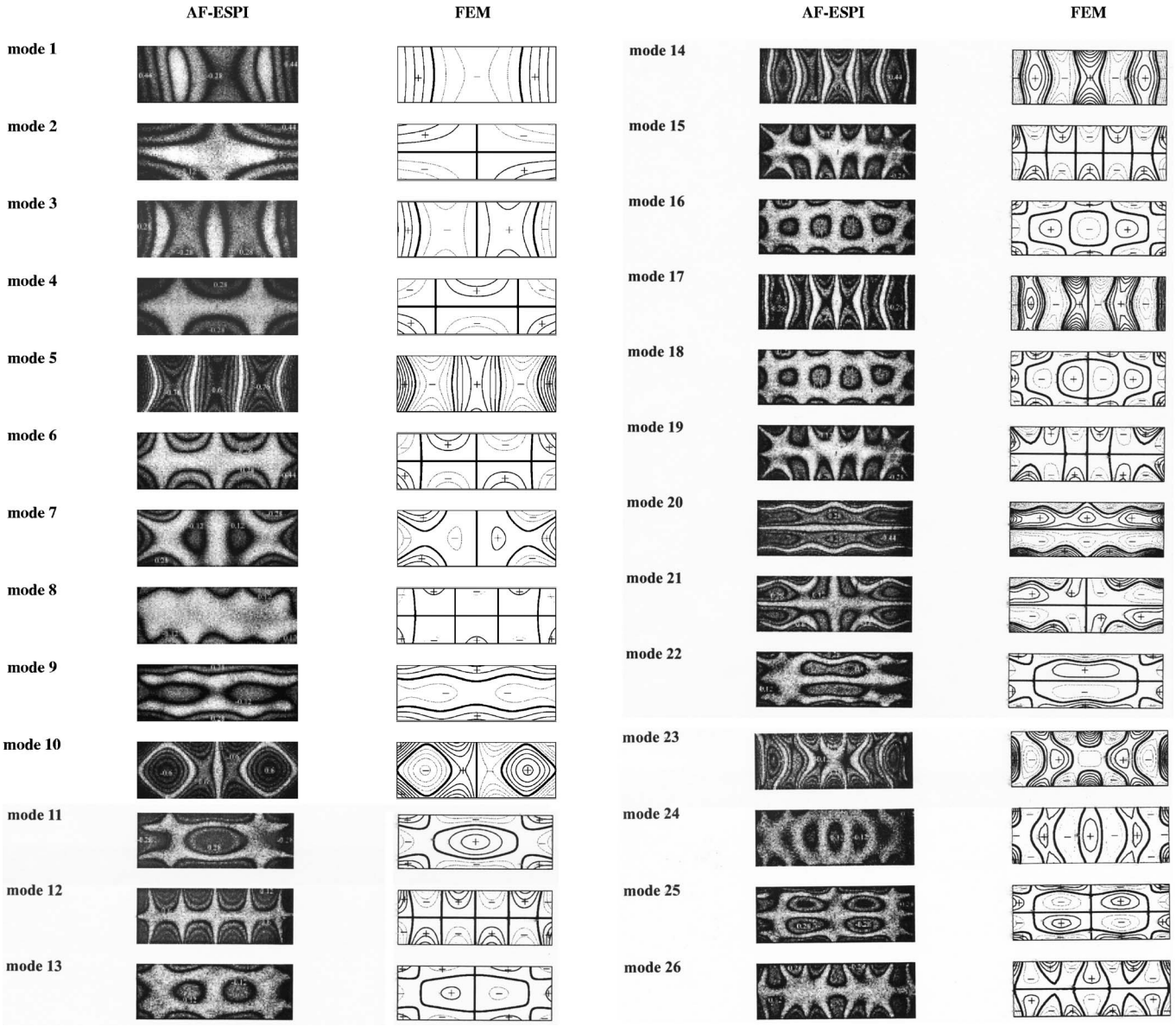


FIG. 4. The out-of-plane mode shapes of the thin piezoceramic plate obtained by AF-ESPI and FEM.

is obtained from the temperature and the isotropic thermal expansion coefficient. In this study, the material parameters of the soft sponge are determined by putting the counterweights on the upper surface of the sponge and measuring the correspondent displacements. The value of N is set to be 1. The material parameters u_i, α_i, and β_i are determined to be 10 303.2, 25.0, and 0, respectively, from the measurement of the nominal stress and strain in a simple experiment. Since we do not take the thermal effect into account, hence $\varepsilon^{\text{th}}=0$, $J^{\text{th}}=1$, and $J^{\text{el}}=\lambda_1\lambda_2\lambda_3$.

Figure 4 presents the detailed mode shapes of the PIC-151 single-layer piezoceramic plate obtained by AF-ESPI with out-of-plane sensitivity and FEM. The excellent quality of the interferometric fringe pattern obtained from the AF-ESPI method is demonstrated. For the finite element results as shown on the right-hand side of Fig. 4, the dashed lines and the "−" sign indicate the concave displacements, while the solid lines and the "+" sign denote convex displacements. The transition from solid lines to dashed lines corresponds to a zero displacement line or a nodal line that is represented as a bold black line. The zero-order fringes, which are the brightest fringes on the experimental results, represent the nodal lines of the piezoceramic plate at resonant frequencies. The rest of the fringes are contour of constant displacements. The mode shapes obtained from experimental results can be checked by the nodal lines and fringe patterns with the numerical finite element calculations. The related amplitude A_i, $i=1,2,3,4\ldots,n$ for the ith fringe in the experimental results can be quantitatively calculated by the roots of $J_0(\Gamma A)=0$ as indicated in Eq. (6). We use $\theta=10°$ for the experimental setup and $\lambda=632.8$ nm; the related vibration amplitudes for the first six dark fringes are A_i, $i=1-6$, $=0.12$, 0.28, 0.44, 0.6, 0.76, and 0.92 μm. The maximum value of the vibration displacement and related amplitudes of some fringes in the experimental results are also indicated in Fig. 4. Note that the vibration displacements obtained in this study are in the order of submicrometer. Excellent agree-

TABLE II. Results of resonant frequencies for the out-of-plane vibration obtained from AF-ESPI and FEM for a thin PIC-151 plate (FEM I represents the complete free case, FEM II represents the sponge effect included, Error is compared with AF-ESPI results).

Mode	AF-ESPI (Hz)	FEM I (Error %)	FEM II (Error %)
1	221	165.3(−33.73)	220.7(−0.12)
2	306	257.4(−18.90)	319.9(4.35)
3	525	466.9(−11.07)	503.5(−4.28)
4	613	556.6(−10.14)	597.3(−2.62)
5	953	923.5(−3.20)	950.0(−0.31)
6	990	936.8(−5.68)	968.6(−2.21)
7	1466	1419.1(−3.30)	1444.5(−1.49)
8	1491	1432.2(−4.11)	1459.2(−2.18)
9	1509	1503.2(−0.39)	1527.8(1.23)
10	1746	1746.7(0.04)	1769.4 (1.32)
11	1896	1859.5(−1.96)	1880.1(−0.85)
12	2127	2066.1(−2.95)	2094.7(−1.54)
13	2401	2345.5(−2.37)	2364.6(−1.54)
14	2486	2471.9(−0.57)	2499.8(0.55)
15	2905	2844.4(−2.13)	2881.9(−0.80)
16	3022	2963.4(−1.98)	2984.8(−1.25)
17	3422	3435.3(0.39)	3481.8(1.72)
18	3785	3705.5(−2.15)	3760.5(−0.65)
19	3799	3723.7(−2.02)	3764.2(−0.92)
20	4097	4104.9(0.19)	4166.3(1.66)
21	4296	4298.5(0.06)	4365.2(1.59)
22	4461	4489.7(0.64)	4544.5(1.84)
23	4529	4521.0(−0.18)	4588.5(1.30)
24	4702	4687.5(−0.31)	4772.8(1.51)
25	4985	5000.8(0.32)	5074.7(1.77)
26	5077	5088.9(0.23)	5194(2.30)

ments of the experimental measurement and numerical calculation are found for all the resonant modes.

Table II tabulates the first 26 resonant frequencies obtained by AF-ESPI and FEM, where FEM I represents the first case, the completely free boundary, FEM II represents the second case that includes the effect of the soft sponge. Compare resonant frequencies measured by the AF-ESPI method with the results predicted by FEM I, the larger discrepancy is greater than 10% and occurs at the first four modes; good agreement is obtained for the higher vibration modes. However, if we compare the resonant frequencies obtained by AF-ESPI and FEM II, excellent agreement is achieved for all the vibration modes and the difference for the first four modes are significantly reduced. It can be concluded at this point that the soft sponge support has a significant influence on the resonant frequencies of the lower resonant modes for thin piezoceramic plates. It is worthy to note that the sequence of all the 26 modes calculated from FEM II is the same as that obtained by the experimental observation. However, the sequence of the 18th and 19th modes are exchanged from the prediction by FEM I. It is indicated from Table II that the finite element simulation underestimates the resonant frequencies if the effect of the sponge support is not included.

It is noted that if the geometrical dimension and material properties are known for the tested specimen and the convergence investigation has been performed, then resonant frequencies can be accurately determined from numerical calculations. However, some difficulties may arise for the AF-ESPI experimental measurement of resonant frequencies. From the experimental measurement point of view, it is well known that there is a bandwidth for the resonant frequency determination. Some resonant frequencies have very narrow bandwidth for only several hertz and the resonant frequency can be easily and accurately determined experimentally. However, some resonant frequencies have relatively wide bandwidth for hundreds of hertz. The number of fringes is nearly the same within the bandwidth of resonant frequencies and the experimental result is not easy to accurately determine. Furthermore, the material properties provided by Physik Instrumente and the boundary condition of the specimen might be different between experimental measurement and numerical calculation. The above-mentioned main reasons will result in the discrepancy of resonant frequencies measured by experiment and predicted by finite element calculation.

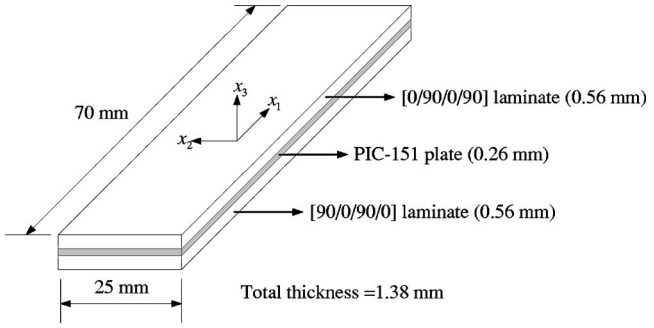


FIG. 5. Geometrical configuration of the GFRP piezolaminated composite plate.

B. The results of the multilayer cross-ply piezolaminated composite plate

The resonant characteristics of in-plane and out-of-plane vibrations for a cross-ply piezolaminated composite plate with an embedded piezoceramic layer are investigated. The optical systems for the out-of-plane and the in-plane measurements are indicated in Figs. 1 and 2, respectively. The piezolaminated composite plate has the size of $70 \times 25 \times 1.38$ mm. Figure 5 shows its schematic illustration of the specimen, a thin piezoceramic plate that has the same size and material property as the first experimental specimen is embedded in the middle of the cross-ply fiberglass reinforced plastic materials. The stacking sequence of the piezolaminated composite plate is [0/90/0/90/PIC-151/90/0/90/0]. The two opposite faces of the thin piezoceramic plate are completely coated with silver electrodes, and this cross-ply piezolaminated composite plate is excited by the application of a time-harmonic voltage across electrodes on the two sur-

TABLE III. Material constants of fiberglass reinforced lamina used in FEM I (unidirectional).

Material properties	E_{11} (GPa)	E_{22} (GPa)	ν_{12}	G_{12} (GPa)	ρ
Magnitude	36.11	8.8	0.32	3.48	1820 (kg/m^3)

TABLE IV. Material constants of fiberglass reinforced lamina used in FEM II and FEM III (unidirectional).

Material properties	E_{11}	E_{22}	E_{33}	ν_{12}	ν_{13}	ν_{23}	G_{12}	G_{13}	G_{23}	ρ
Magnitude	36.11 (GPa)	8.8 (GPa)	8.8 (GPa)	0.32	0.32	0.43	3.48 (GPa)	3.48 (GPa)	3.1 (GPa)	1820 (kg/m^3)

faces of the embedded thin piezoceramic plate. The fiberglass used is a continuous filament "E" type, the composite prepreg tape is manufactured by the Forthtrack Co., Ltd, and the model number is FT-150G. It consists of fiberglass precoated with the thermosetting resin, the composition of epoxy resin and fiberglass is 37% and 63%, respectively. The fabrication of these laminated composite structures involves simply laying-up the prepreg at the required orientation on a forming mold, stacking layers of prepreg in the required stacking sequence, and then curing the assembly under elevated temperature and pressure of 130 °C and 50 psi for 2 h. Table III lists the material properties of the fiberglass-reinforced lamina measured by the material testing system (MTS) and strain gauges, where E_{11} and E_{22} are Young's moduli in the longitudinal and transverse directions, respectively; ν_{12} is the Poisson's ratio, G_{12} is the in-plane shear modulus. The material properties listed in Table III are enough for the two-dimensional FEM calculation. However, more material constants of the fiberglass-reinforced lamina are needed for the complicated three-dimensional numerical calculation. Since the axial axis of the fiber in the fiberglass-reinforced composite is along the x_1 axis, it is often assumed that the material behavior in the x_2 axis is identical to the material behavior in the x_3 axis. For this situation, we have $E_{33}=E_{22}=8.8$ GPa, $\nu_{13}=\nu_{12}=0.32$, and $G_{13}=G_{12}=3.48$ GPa. We further assume that the material is isotropic in the x_2-x_3 plane and $\nu_{23}=0.43$. The shear modulus can then be evaluated from $G_{23}=E_2/2(1+\nu_{23})=3.1$ GPa. Hence the other five elastic constants $(E_{33},\nu_{13},\nu_{23},G_{13},G_{23})$ are properly chosen without using the experimental measurement and the complete list of the material properties is indicated in Table IV. Because the piezolaminated composite plate has the thickness of 1.38 cm, both the in-plane and out-of-plane vibration modes will be discussed. The optical setups of AF-ESPI for the in-plane and out-of-plane sensitivity are used to investigate the resonant characteristics of the piezolaminated composite plate. The vibration analysis of angle-ply laminated composite plates with an embedded piezoceramic layer was investigated by Lin *et al.*[16] They analyzed the resonant frequencies and correspondent mode shapes of five different angle-plies for out-of-plane modes only.

In FEM simulation for the multilayer cross-ply piezolaminated composite plate, three formulations based on different models are used to evaluate the resonant vibration characteristics of the piezolaminated plate. The Lanczos eigensolver is used to perform eigenvalue extraction for calculating the natural frequencies and corresponding mode shapes. The numerical results for resonant frequencies are listed in Tables V and VI and are denoted as FEM I, FEM II, and FEM III. FEM I uses the plane shell element (S9R5) with nine-node, five degrees of freedom at the node, and the Kirchhoff constraint imposed numerically is used to model the resonant vibration behaviors of the piezolaminated plate. The material properties used in the numerical calculation for FEM I are indicated in Tables I and III. A local coordinate system for definition of the material properties in each layer is used. FEM II uses the solid element (C3D20R) to model the resonant vibration behavior of the piezolaminated plate. The total numbers of elements and nodes used are 2646 and 13 974, respectively. Because the piezoelectric coupling effect is active only in piezoelectric elements, the effect of the embedded piezoceramic layer on the resonant vibration characteristics is not included in FEM I and FEM II. To investigate the piezoelectric coupling effect of the embedded piezoceramic layer, the solid element (C3D20R) and the three-dimensional piezoelectric element (C3D20RE) are used in FEM III to model the behavior of laminated composite layers and the embedded piezoceramic layer, respectively. The elements C3D20R and C3D20RE are 20-node quadratic brick, reduced integration, three-dimensional continuum stress/displacement and piezoelectric element, respectively. The total numbers of elements and nodes used in FEM III are the

TABLE V. Results of resonant frequencies for the out-of-plane vibration obtained from AF-ESPI, LDV, and FEM for the piezolaminated composite plate. (Error is compared with AF-ESPI values.)

Mode	AF-ESPI (Hz)	LDV (Error %)	FEM I (Error %)	FEM II (Error %)	FEM III (Error %)
1	976	980(0.41)	885(−9.32)	806(−17.39)	809(−17.12)
2	1005	1030(2.49)	918(−8.66)	889(−11.56)	889(−11.56)
3	2185	2240(2.52)	2082(−4.71)	1787(−9.06)	1989(−9.00)
4	2605	2600(−0.19)	2440(−6.33)	2215(−14.96)	2223(−14.66)
5	3838	3970(3.44)	3744(−2.45)	3508(−8.60)	3514(−8.43)
6	4964	5050(1.73)	4778(−3.75)	4316(−13.06)	4332(−12.74)
7	6197	6070(−2.05)	5847(−5.65)	5602(−9.60)	5617(−9.37)
8	6362	6360(−0.03)	6083(−4.39)	6112(−3.94)	6140(−3.50)
9	6697	6600(−1.45)	6113(−8.72)	6327(−5.53)	6350(−5.18)
10	7338	7510(2.34)	7028(−4.22)	7117(−3.02)	7149(−2.58)

TABLE VI. Convergence test of out-of-plane resonant frequencies from FEM III of the piezolaminated composite plate.

No. of elements	120	210	324	552	702	1188	1638	2112	2646	3078
Mode 1	808.3	808.5	808.7	808.8	808.8	808.9	808.9	808.9	808.9	808.9
Mode 2	895.1	893.5	892.4	891.0	890.5	889.6	889.3	889.0	888.9	888.8
Mode 3	2000.2	1997.2	1995.1	1992.6	1991.8	1990.2	1989.6	1989.1	1988.8	1988.7
Mode 4	2221.9	2221.8	2222.1	2222.5	2222.6	2222.8	2222.9	2222.9	2223	2223
Mode 5	3529.3	3524.3	3521.6	3518.7	3517.8	3516	3515.2	3514.7	3514.3	3514.1
Mode 6	4336.9	4330.8	4330.2	4330.9	4331	4331.4	4331.6	4331.6	4331.7	4331.7
Mode 7	5637.2	5626.2	5622.8	5620.4	5619.6	5618.1	5617.4	5616.9	5616.6	5616.4
Mode 8	6249.4	6204.4	6182.7	6162	6156.3	6146.7	6143.3	6141.1	6139.6	6139
Mode 9	6454.2	6412.6	6392.4	6372.7	6367.1	6357.4	6353.9	6351.6	6350	6349.4
Mode 10	7190.8	7156.5	7149.8	7148.6	7148.4	7148.5	7148.5	7148.5	7148.6	7148.6

same as the case in FEM II. For the case of FEM III, the electrical boundary condition is set to be short-circuited on both the top and bottom surfaces. It is obvious that FEM II and FEM III are more time consuming for the numerical calculation. The material properties used in the numerical calculation for FEM II and FEM III are indicated in Tables I and IV.

Figures 6 and 7 show the out-of-plane and in-plane resonant mode shapes obtained by AF-ESPI and FEM I, respectively. The first ten transverse (out-of-plane) modes of the piezolaminated composite plate are indicated in Fig. 6. It is found that the sequence of the third and the fourth, the fifth and the sixth, the seventh and the ninth, the tenth and the eleventh mode shapes of the piezolaminated composite plate interchanges as compared with the mode shapes of the single-layer piezoceramic plate as shown in Fig. 4. The sequence and mode shapes for out-of plane vibration modes predicted by FEM I, FEM II, and FEM III are almost the same, only the sequence of the seventh and the eighth modes are exchanged for FEM I and FEM II (FEM III). Owing to the geometric configuration of the rectangular piezolaminated composite plate, two in-plane vibration modes for each resonant frequency in the x_1 and x_2 directions are displayed in order to present the complete characteristics of the in-plane vibration behavior. Figure 7 shows the results of the first six in-plane mode shapes, the U field and the V field denote the full-field vibration displacement along the x_1 and the x_2 directions, respectively. The V field of the sixth mode cannot be measured by the AF-ESPI, this is due to the reason that the vibration amplitude for this mode is too small and is beyond the sensitivity of the in-plane setup of the AF-ESPI. It is shown in Fig. 7 that the characteristics of the vibration displacements for the U and V fields are quite different. Excellent agreement of the experimental measurement and numerical calculation is found for both the out-of-plane and in-plane vibration modes as indicated in Figs. 6 and 7.

Because the electrical impedance will drop to a local minimum when it is in resonance, the resonant frequencies can also be obtained by the impedance analysis. Figure 8 shows the impedance curve of this piezolaminated composite plate measured by a HP4194A impedance/gain-phase analyzer. The local minimum appearing in the impedance curve corresponds to the resonant frequencies. Because the summation of the induced charge distributed over the electrode surfaces is zero for out-of-plane modes, only the in-plane modes can be measured by the impedance analysis.

The LDV is a pointwise displacement measurement technique and is sensitive to the out-of-plane displacement. Hence, the LDV is used as the second experimental technique to verify the experimental results of the transverse resonant frequencies for the piezolaminated composite plate obtained by AF-ESPI. The optical system of LDV is based on the principle of Michelson interferometer and the Doppler effect. A built-in dynamic signal analyzer (DSA) is integrated into the LDV system to become the LDV-DSA sys-

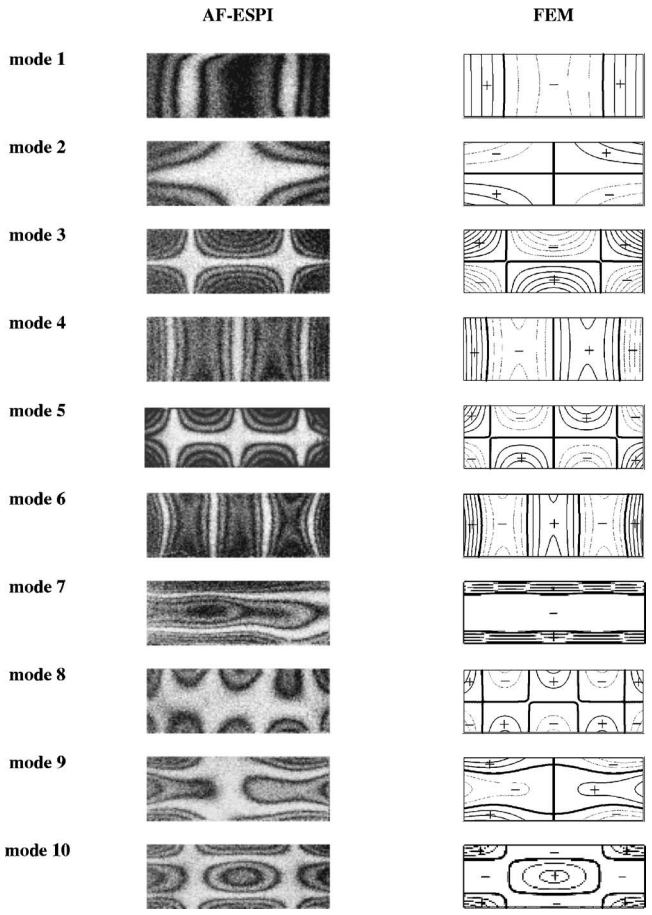


FIG. 6. The out-of-plane mode shapes of the GFRP piezolaminated composite plate obtained by AF-ESPI and FEM.

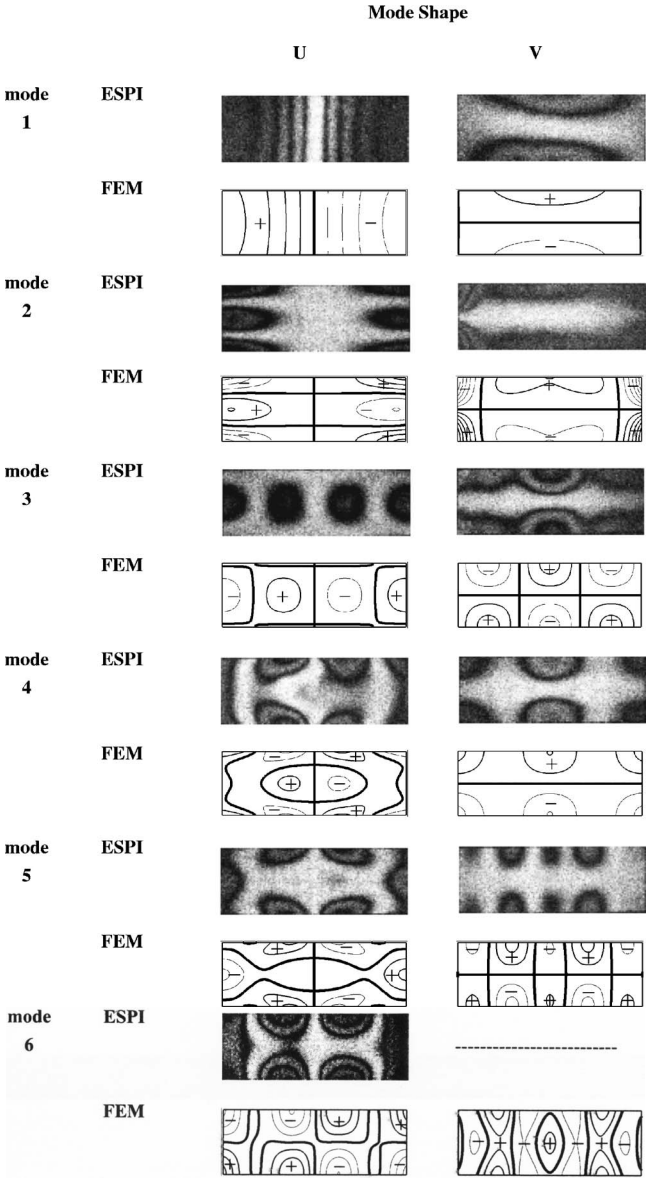


FIG. 7. The in-plane mode shapes of the GFRP piezolaminated composite plate obtained by AF-ESPI and FEM.

tem. The DSA unit is composed of an analysis software and a plug-in wave form generator board that can provide the swept-sine excitation signal. In the analysis software, the swept-sine excitation signal is taken as input and the response measured by LDV is conversed into the voltage signal and is taken as output. After the fast Fourier transform (FFT) of the input and output with the DSA software, the ratio of output/input ("gain") is obtained. The result chart that shows the frequency response curve can be obtained. The peak values appearing in the frequency response curve are the resonant frequencies for the out-of-plane vibration modes. Figure 9 indicates the frequency response curve of the piezolaminated composite plate measured by the LDV-DSA system. A detailed description of LDV-DSA system and the working principle can be found in Ref. 17.

Table V lists the resonant frequencies obtained by AF-ESPI, LDV, and FEM (FEM I, FEM II, FEM III) for out-of-plane modes. The largest discrepancy occurs in mode 1 for

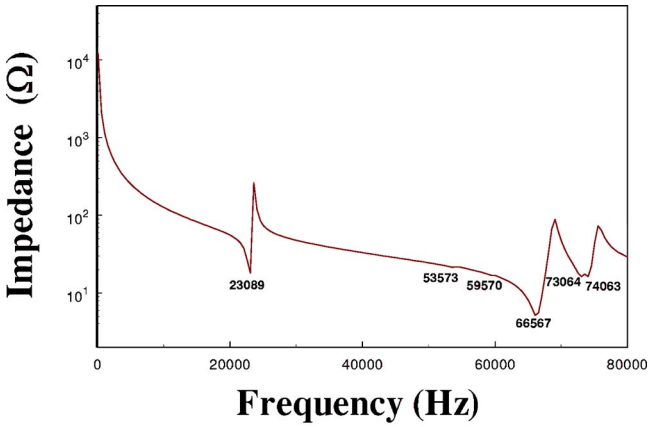


FIG. 8. Impedance curve of the piezolaminated composite plate.

FEM when compared with the AF-ESPI results. We can see that the resonant frequencies measured from these two optical methods agree very well. The resonant frequencies of out-of-plane modes predicted by FEM II and FEM III (three-dimensional analysis) are almost the same. The two experimental measured resonant frequencies (AF-ESPI and LDV) are closely related to the predicted results by FEM I (two-dimensional analysis), but have larger discrepancies when compared with the results by FEM II and FEM III. This is due to the fact that five of the material properties used for the numerical calculations by FEM II and FEM III are not directly obtained from the experimental measurement. The convergence test of the finite element result FEM III has been performed and the result is indicated in Table VI. It shows that the numerical result of the resonant frequency has no difference as the element number over 2000. It is indicated previously that the total number of element used for FEM II an FEM III is 2646.

Table VII shows the first six resonant frequencies obtained by AF-ESPI, impedance, and FEM for in-plane vibration modes. It is worthy to note that the results obtained by impedance analyzer, AF-ESPI and FEM I are in excellent agreement. The largest discrepancy is only 1.57%. The discrepancy of the results between experimental measurement

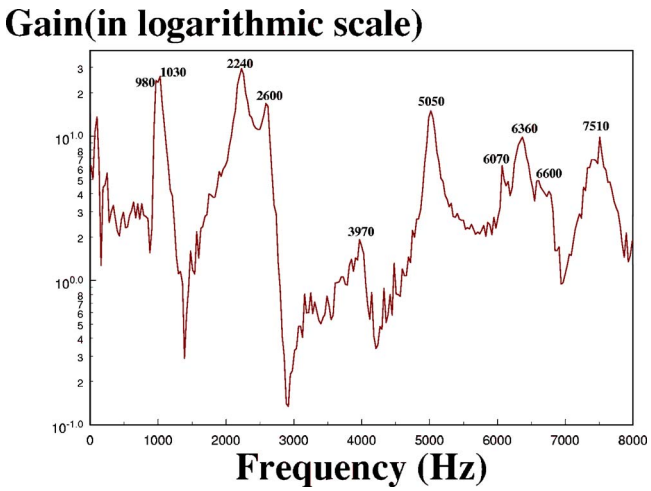


FIG. 9. Frequency response of the piezolaminated composite plate obtained by LDV-DSA system.

TABLE VII. Results of resonant frequencies for the in-plane vibration obtained from AF-ESPI, impedance and FEM for the piezolaminated composite plate. (Error is compared with AF-ESPI values.)

Mode	AF-ESPI (Hz)	Impedance (Error %)	FEM I (Error %)	FEM II (Error %)	FEM III (Error %)
1	22 975	23 089(0.50)	22 924(−0.22)	22 713(−1.14)	23 660(2.98)
2	53 450	53 573(0.23)	53 162(−0.54)	51 553(−3.55)	52 562(−1.66)
3	59 400	59 570(0.29)	60 331(1.57)	58 745(−1.10)	58 842(−0.94)
4	66 240	66 567(0.49)	65 767(−0.71)	64 426(−2.74)	68 611(3.58)
5	72 800	73 064(0.36)	72 745(−0.08)	70 652(−2.95)	73 075(0.38)
6	74 060	74 063(0.00)	74 563(0.68)	72 920(−1.54)	74 353(0.40)

and FEM II (FEM III) is slightly larger than that between experimental measurement and FEM I. The material constants of the composite laminate used for FEM II (FEM III) are only nominal, therefore there might be differences between them and actual constants used in the experiments. Compare the results obtained by FEM II and FEM III, it is interesting to note that most of the results for resonant frequencies predicted by the FEM III are larger than those by FEM II. It is found that the piezoelectric effect only leads to a slighter increase in the rigidity of piezolaminated plates. The influence of the electric field on the free vibrations of piezoelectric materials was also investigated by Krommer and Irschik[18] and Borrelli and Patria.[19]

IV. CONCLUSIONS

It is known that the resonant characteristics of piezoceramic plates are important in many engineering applications. However, there are only very few experimental results available in the literature. This study investigates the resonant frequencies and mode shapes of piezolaminated composite plate for out-of-plane and in-plane vibrations by three experimental techniques (AF-ESPI, LDV, impedance) and FEM. It has been shown that the AF-ESPI method has the advantages of noncontact, full-field, real-time, and high-resolution measurement. Excellent quality of the interferometric fringe for the mode shape is presented by a video recording system. For the thin piezoceramic plate, the resonant frequencies and full-field mode shapes for up to 26 modes are presented by AF-ESPI and are excellently correlated with FEM results. It is found that the soft sponge support has a significant influence on the resonant frequencies of the lower resonant modes for thin piezoceramic plates. The finite element simulation underestimates the resonant frequencies for thin piezoceramic plates if the effect of the sponge support is not included. For the case of the piezolaminated composite plate, not only transverse modes but also in-plane modes are analyzed by three experimental techniques and FEM. Excellent agreement between the theoretical predictions and experimental measurements of resonant frequencies and mode shapes is obtained. The electric boundary conditions applied on the piezoceramic plate of the piezolaminated composite specimen have nearly no influence on the resonant frequencies for the out-of-plane modes, but significant influences are found for the in-plane vibration modes. The results indicated in this study also demonstrate that the experimental techniques proposed herein are applicable in the resonant analysis for piezolaminated plates.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support of this research by the National Science Council (Republic of China) under Grant No. NSC 90-2212-E-231-003.

[1] S. K. Ha, C. Keilers, and F. K. Chang, "Finite element analysis of composite structures containing distributed piezoceramic sensors and actuators," AIAA J. **30**, 772–780 (1992).
[2] R. C. Batra and X. Q. Liang, "Vibration of a rectangular laminated elastic plate with embedded piezoelectric sensors and actuators," Comput. Struct. **63**, 203–216 (1997).
[3] H. Abramovich and H. R. Meyer-Piening, "Induced vibrations of piezolaminated elastic beams," Compos. Struct. **43**, 47–55 (1998).
[4] M. A. R. Loja, J. I. Barbosa, C. M. Mota Soares, and C. A. Mota Soares, "Analysis of piezolaminated plates by the spline finite strip method," Comput. Struct. **79**, 2321–2333 (2001).
[5] J. N. Butters and J. A. Leendertz, "Speckle patterns and holographic techniques in engineering metrology," Opt. Laser Technol. **3**, 26–30 (1971).
[6] R. Jones and C. Wykes, *Holographic and Speckle Interferometry* (Cambridge University Press, New York, 1982).
[7] K. Hogmoen and O. J. Løkberg, "Use of modulated reference wave in electronic speckle pattern interferometry," J. Phys. E **9**, 847–851 (1976).
[8] S. Nakadate, "Vibration measurement using phase-shifting speckle pattern interferometry," Appl. Opt. **25**, 4162–4167 (1986).
[9] J. D. Valera, A. F. Doval, and J. D. C. Jones, "Determination of vibration phase with electronic speckle pattern interferometry (ESPI)," Electron. Lett. **28**, 2292–2294 (1992).
[10] W. O. Wong, K. T. Chan, and T. P. Leung, "Contrast and sensitivity of the vibration fringes in time-averaged electronic speckle-pattern interferometry: Effect of variations of force level," Opt. Laser Technol. **29**, 179–185 (1997).
[11] W. C. Wang, C. H. Huang, and S. Y. Lin, "Vibration measurement by the time-averaging electronic speckle pattern interferometry method," Appl. Opt. **35**, 4502–4509 (1996).
[12] W. C. Wang and Y. H. Tsai, "Experimental vibration analysis of the shadow mask," Opt. Lasers Eng. **30**, 539–550 (1998).
[13] C. C. Ma and C. H. Huang, "The investigation of three-dimensional vibration for piezoelectric rectangular parallelepipeds using the AF-ESPI method," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **48**, 142–153 (2001).
[14] C. H. Huang and C. C. Ma, "Vibration characteristics for piezoelectric cylinders using amplitude-fluctuation electronic speckle pattern interferometry," AIAA J. **36**, 2262–2268 (1998).
[15] *ABAQUS User's Manual, ver. 6.2*, Hibbit, Karlsson, and Sorensen, Inc, Pawtucket, RI, 2001.
[16] H. Y. Lin, J. H. Huang, and C. C. Ma, "Vibration analysis of angle-ply laminated composite plates with an embedded piezoceramic layer," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **50**, 1084–1099 (2003).
[17] C. K. Lee and G. Y. Wu, "High performance Doppler interferometer for advanced optical storage system," Jpn. J. Appl. Phys., Part 1 **38**, 1730–1741 (1999).
[18] M. Krommer and H. Irschik, "On the influence of the electric field on free transverse vibrations of smart beams," Smart Mater. Struct. **8**, 401–410 (1999).
[19] A. Borrelli and M. C. Patria, "Saint-Venant's principle for a piezoelectric body," SIAM J. Appl. Math. **59**, 1098–1111 (1999).

Centralized and decentralized control of structural vibration and sound radiation

Wouter P. Engels,[a] Oliver N. Baumann, and Stephen J. Elliott
Institute of Sound and Vibration Research, University of Southampton, University Road, Highfield, Southampton, SO17 1BJ, United Kingdom

R. Fraanje
Delft Center for Systems and Control, Delft University, Mekelweg 2, 2628 CD, Delft, The Netherlands

(Received 23 March 2005; revised 15 November 2005; accepted 11 December 2005)

This paper examines the performance of centralized and decentralized feedback controllers on a plate with multiple colocated velocity sensors and force actuators. The performance is measured by the reduction in either kinetic energy or sound radiation, when the plate is excited with a randomly distributed, white pressure field or colored noise. The trade-off between performance and control effort is examined for each case. The controllers examined are decentralized absolute velocity feedback, centralized absolute velocity feedback control and linear quadratic Gaussian (LQG) control. It is seen that, despite the fact that LQG control is a centralized, dynamic controller, there is little overall performance improvement in comparison to decentralized direct velocity feedback control if both are limited to the same control effort. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2163270]

PACS number(s): 43.40.Vn [KAC] Pages: 1487–1495

I. INTRODUCTION

Various control strategies can be used to control the vibration of plates. They can be aimed specifically at controlling the kinetic energy of the plate (active vibration control, AVC) or the sound radiation (active structural acoustic control, ASAC). If a reference signal is not available, control strategies are limited to the use of feedback controllers. These can vary greatly in complexity. The complexity of the controller is understood here to be determined by both the number of states in the controller and whether the controller is centralized or decentralized.

Decentralized, static gain control is the simplest form of feedback control. If it is applied in a stable system where the sensors and actuators are colocated and dual, then stability is, in theory, guaranteed [Balas (1979), Sun (1996)]. In a practical situation, it can have the extra advantage that no connections are required between different control locations and/or a central processing unit and that actuator, sensor and controller could be produced as identical modular units. Decentralized feedback control has been examined by, amongst others Elliott *et al.* (2002), who compare the performance using colocated force actuators and velocity sensors with piezoelectric actuators and velocity sensors. Gardonio *et al.* (2004) gives an extensive review of control methods in ASAC as well as examining the total sound radiation of a panel as a function of a centrally set feedback gain for 16 control loops consisting of piezoceramic patches as actuators and accelerometers as sensors. This means that sensor and actuator are not dual and the stability of the feedback loop is limited. Also, by limiting the system to have the same gain in each control loop, the feedback gains are not optimal. Engels *et al.* (2004) compares centralized and decentralized, constant gain control on a beam, but does not take the control effort into account.

More complex controllers have also been examined. Fuller *et al.* (2004) describes heterogeneous blankets for AVC and ASAC, that essentially consist of numerous mass-spring resonators. These resonators could be viewed as decentralized, colocated, dynamic feedback loops. Bingham *et al.* (2001) examines different strategies in several single-input–single-output loops applied to the same plate and found that more complex strategies do result in better performance, but the control loops were not dual, nor was control effort examined. Clark and Cox (1997) compares LQG control and a centralized constant gain controller, optimized for ASAC on a plate with dual control loops and showed that constant gain, velocity feedback can be an effective alternative for dynamic controllers. Though the control effort weighting was the same in the cost function minimized by both control strategies, the actual effort may still have been considerably different. Following-up on this work, Smith and Clark (1998) compared the acoustic performance of static feedback gains in different controller arrangements. Each controller was used to minimize specific cost functions, but now the control effort weighting was adjusted such that the different controllers used the same control effort. For a single channel controller, LQG control and static feedback control were also compared. It was shown that designing for ASAC did result in slightly improved sound reductions for all controllers and that, for a single channel case with a large, distributed sensoriactuator, LQG control gave better performance than static feedback control. However, the complexity of the LQG controller makes it considerably more difficult to implement.

[a]Electronic mail: we@isvr.soton.ac.uk

TABLE I. Variables of the plate used in the simulations.

$E=7\times10^{10}$ (Pa)	$\rho=2720$ (kg/m^3)
$\zeta=0.01$	$h=0.001$ (m)
$l_x=0.247$ (m)	$l_y=0.278$ (m)
$f_{\max}=3$ (kHz)	$I=h^3/12$ (m^3)

To select an appropriate degree of complexity for the controller, the performance of the different strategies needs to be compared on an even footing. In particular, the control effort that is applied by the controller must be taken into account, as optimization of the cost function for different controllers tends to result in different control efforts. It is also important to examine the trade-off between performance and control effort for different controller designs to select an appropriate amount of control effort. This paper will compare the performance and control effort for a decentralized constant gain controller, a centralized constant gain controller and a centralized linear quadratic Gaussian (LQG) controller. The controllers will be optimized for different performance measures and different excitation spectra of a simply supported plate. The different performance measures are the kinetic energy of the plate or the sound power radiated into the far field. The parameters of the plate used in the study are listed in Table I and it is assumed to be excited by a spatially completely random pressure field with a spectrum corresponding to white noise. This is an idealization of a turbulent boundary layer excitation, as in reality the spatial correlation of turbulent boundary layer has some finite value [Maury *et al.* (2002)]. Colored noise is also examined. Sixteen equally spaced, colocated actuators and sensors are assumed to be arranged on the plate, as shown in Fig. 1. The actuators are assumed to be ideal point force actuators and the sensors ideal velocity sensors.

II. STRUCTURAL MODEL

The model of the simply supported plate, consists of a modal model as described in Fuller *et al.*(1996). The mode shapes are defined as

$$\Psi_{mn}(x,y)=\sin(k_m x)\sin(k_n y) \tag{1}$$

with $k_m=m\pi/l_x$ and $k_n=n\pi/l_y$. The natural frequency corresponding to the above mode is

$$\omega_{mn}=\sqrt{(EI/\rho h)}(k_m^2+k_n^2). \tag{2}$$

A small amount of damping, ζ is included in the response of the modes. The response of the amplitude of each mode, a_{mn}, to a harmonic point force, $F_\omega(j\omega)$, acting at a point (x,y) on the plate then is

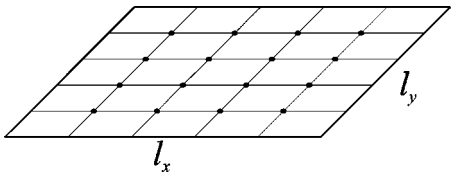


FIG. 1. Equally spaced sensors and actuators. Each dot represents colocated velocity sensors and point force actuator pair.

$$a_{mn}(j\omega)=\frac{4\Psi_{mn}(x,y)F_\omega(j\omega)}{M(\omega_{mn}^2+2j\zeta\omega\omega_{mn}-\omega^2)}, \tag{3}$$

where M is the total mass of the plate. If only a finite number, N, of modes is taken into account and the total response of the system can be written in a state space model,

$$\begin{pmatrix}\dot{\mathbf{a}}_s\\ \ddot{\mathbf{a}}_s\end{pmatrix}=\begin{bmatrix}\mathbf{0} & \mathbf{I}\\ -\mathbf{K}_s & -\mathbf{D}_s\end{bmatrix}\begin{pmatrix}\mathbf{a}_s\\ \dot{\mathbf{a}}_s\end{pmatrix}+\frac{4}{M}\begin{pmatrix}\mathbf{0}\\ \mathbf{\Psi}\end{pmatrix}\mathbf{F}_t(t), \tag{4}$$

where $\mathbf{a}_s$ and $\dot{\mathbf{a}}_s$ are, respectively, vectors of the modal amplitude and modal velocity. The mode shapes taken into account are all the mode shapes with a natural frequency up to $f_{\max}$, which results in 60 modes taken into account. The matrix $\mathbf{K}_s$ is a matrix with the squared natural frequencies, ω_{mn}^2, on its diagonal and empty otherwise. The matrix $\mathbf{D}_s$ is also a diagonal matrix, but with the damping terms of the denominator of Eq. (3), $2\zeta\omega_{mn}$ on its diagonal.

For the chosen mode shapes, the kinetic energy of the plate at any particular point in time, can be calculated as the sum of the squared modal velocities [Meirovitch (1986)]

$$J_{\text{ke}}=\frac{M}{8}\dot{\mathbf{a}}_s^T(t)\dot{\mathbf{a}}_s(t). \tag{5}$$

III. RADIATION MODEL

The modeling of sound radiation of a plate in an infinite baffle is usually done in one of two ways, either by analyzing modal radiation or the so-called radiation modes. Modal radiation models the autoradiation and cross radiation of the structural modes and has been used in several papers [Baumann *et al.* (1991), Thomas and Nelson (1995), Clark and Frampton (1999), Vipperman and Clark (1991)] to examine and implement ASAC. The radiation modes are velocity distributions of the structure surface that radiate sound independently and are based on the work by Borgiotti (1990), Elliott and Johnson (1993), and Borgiotti and Jones (1994). These velocity distributions vary with frequency but only slowly so. By assuming that the velocity distributions do not change with frequency, an approximation of the radiated sound power can be obtained. This approach may be called the fixed radiation mode approach [Cox *et al.* (1998), Gibbs *et al.* (2000), Gardonio and Elliott (2004)]. Gibbs *et al.* (2000) called this method radiation modal expansion (RME). Though this approach is limited in the frequency range for which it can accurately model the sound radiation, it requires far less states than the full modal radiation approach, used by Baumann *et al.* (1991), as for RME the required number of filters is equal to the number of radiation modes, rather than proportional to the square of the number of structural modes for the modal radiation approach.

In RME, a set of the most significantly radiating modes is chosen at a specific base frequency. Then, the radiated sound power of these modes is calculated at other frequencies and filters are fitted whose squared response matches these values. By calculating the input to each radiation mode as a function of the modal velocities and then applying the frequency dependent filters, the sound radiation J_{ac} can be estimated as

$$\dot{\mathbf{a}}_{\mathrm{ac}} = \mathbf{A}_{\mathrm{ac}}\mathbf{a}_{\mathrm{ac}} + \mathbf{B}_{\mathrm{ac}}\dot{\mathbf{a}}_s,$$

$$\mathbf{q} = \mathbf{C}_{\mathrm{ac}}\mathbf{a}_{\mathrm{ac}} + \mathbf{D}_{\mathrm{ac}}\dot{\mathbf{a}}_s,$$

$$J_{\mathrm{ac}} = \mathbf{q}^T(t)\mathbf{q}(t), \tag{6}$$

where $\mathbf{a}_{\mathrm{ac}}$ are the states of the radiation filters, the matrix $\mathbf{B}_{\mathrm{ac}}$ describes the excitation of the filters as a function of the modal velocities, and $\mathbf{A}_{\mathrm{ac}}$ describes the dynamics. $\mathbf{C}_{\mathrm{ac}}$ describes the relation between the states of each filter and its output and $\mathbf{D}_{\mathrm{ac}}$ is a direct feedthrough matrix of the structural velocities to the cost variables.

Here, the 20 most significant radiation modes of the plate at 1 kHz have been selected to model the radiation. The inclusion of these filters in the state space model of Eq. (4) leads to a modified state space model,

$$\begin{pmatrix} \dot{\mathbf{a}}_s \\ \ddot{\mathbf{a}}_s \\ \dot{\mathbf{a}}_{\mathrm{ac}} \end{pmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{I} & \mathbf{0} \\ -\mathbf{K}_s & -\mathbf{D}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{\mathrm{ac}} & \mathbf{A}_{\mathrm{ac}} \end{bmatrix} \begin{pmatrix} \mathbf{a}_s \\ \dot{\mathbf{a}}_s \\ \mathbf{a}_{\mathrm{ac}} \end{pmatrix} + \frac{4}{M} \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\Psi} \\ \mathbf{0} \end{pmatrix} \mathbf{F}_t(t),$$

$$\mathbf{q} = [\mathbf{0}\ \mathbf{D}_{\mathrm{ac}}\ \mathbf{C}_{\mathrm{ac}}] \begin{pmatrix} \mathbf{a}_s \\ \dot{\mathbf{a}}_s \\ \mathbf{a}_{\mathrm{ac}} \end{pmatrix}. \tag{7}$$

IV. COST FUNCTION FORMULATION

The two goals of control are to minimize the kinetic energy of the system or the radiated sound power. Also, the control effort used to control the system should be limited. The excitation of the system is assumed to be a random signal in the time domain. Hence, the cost function is chosen to be the time average or expectation, of the two values mentioned above and the control effort, thus resulting in the following expressions for the cost functions relating to kinetic energy, J_{ke}, and sound radiation, J_{ac}, respectively,

$$J_{\mathrm{ke}} = E[M/8\dot{\mathbf{a}}_s^T(t)\dot{\mathbf{a}}_s(t) + \rho\mathbf{u}^T(t)\mathbf{u}(t)], \tag{8}$$

$$J_{\mathrm{ac}} = E[\mathbf{q}^T(t)\mathbf{q}(t) + \rho\mathbf{u}^T(t)\mathbf{u}(t)], \tag{9}$$

where $\mathbf{u}$ is the vector of control signals applied by the controller and ρ is a weighting value of the control effort. To compare the controllers fairly, ρ should be tuned such that the control effort $\mathbf{u}^T(t)\mathbf{u}(t)$ is equal for each controller. Because the variables in the above equations are all dependent on the random excitation of the system, it is useful to rewrite them in terms of the expectation of the excitation.

Both of the above equations can be written in a more general form,

$$J = E[\mathbf{x}^T(t)\mathbf{Q}\mathbf{x}(t) + \mathbf{u}^T(t)\mathbf{R}\mathbf{u}(t)], \tag{10}$$

where $\mathbf{x}$ are the states of the model, in this case, $\mathbf{a}_s$, $\dot{\mathbf{a}}_s$ and, if needed, $\mathbf{a}_{\mathrm{ac}}$. Equations (4) and (7) can then be written in a general state-space form:

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}_u\mathbf{u}(t) + \mathbf{B}_d\mathbf{F}_t(t), \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t), \end{aligned} \tag{11}$$

where the forcing term $\mathbf{F}_t(t)$ has been split into a disturbance force $\mathbf{F}_d$ and a control force $\mathbf{u}$. The matrices $\mathbf{B}_d$ and $\mathbf{B}_u$ are structured in the same way as the matrix preceding the forcing term $\mathbf{F}_t$ in Eqs. (4) and (7).

If a feedback controller is used, the feedback gain is a function of the states of the system and the states of the controller. The terms $\mathbf{u}(t)\mathbf{R}\mathbf{u}(t)$ and $\mathbf{B}_u\mathbf{u}(t)$ can then be removed from Eqs. (10) and (11) without loss of generality, by including the controller in the states $\mathbf{x}(t)$ and the matrix $\mathbf{A}$. The cost of the control effort can then be included in the term $\mathbf{x}^T(t)\mathbf{Q}\mathbf{x}(t)$. To distinguish between the dynamics of the controlled and uncontrolled system, the notation $\mathbf{A}_c$ is used for the controlled system.

The states at any particular time are a convolution of the matrix of impulse responses of the states to a change in any of the states and the excitation of the modes by the disturbance,

$$\mathbf{x}(t) = \boldsymbol{\Phi}(t) * \mathbf{B}_d\mathbf{F}_{t'}(t), \tag{12}$$

where $\boldsymbol{\Phi}(t)$ is the matrix of the impulse responses of the states. This matrix is also known as the *fundamental transition matrix* and is equal to

$$\boldsymbol{\Phi}(t) = e^{\mathbf{A}_c t}. \tag{13}$$

Combining Eqs. (10) and (12) and including control effort in the matrix $\mathbf{Q}$, results in

$$\begin{aligned} J &= E[\mathbf{x}^T(t)\mathbf{Q}\mathbf{x}(t)] = \mathrm{trace}(\mathbf{Q}E[\mathbf{x}(t)\mathbf{x}^T(t)]) \\ &= \mathrm{trace}(\mathbf{Q}E[\boldsymbol{\Phi}(t) * \mathbf{B}_d\mathbf{F}_t(t)(\boldsymbol{\Phi}(t) * \mathbf{B}_d\mathbf{F}_t(t))^T]) \\ &= \mathrm{trace}\Bigg(\mathbf{Q}E\Bigg[\int_0^\infty \boldsymbol{\Phi}(\sigma_1)\mathbf{B}_d\mathbf{F}_t(t-\sigma_1)\mathrm{d}\sigma_1 \\ &\quad \times \int_0^\infty \mathbf{F}_t^T(t-\sigma_2)\mathbf{B}_d^T\boldsymbol{\Phi}^T(\sigma_2)\mathrm{d}\sigma_2\Bigg]\Bigg) \\ &= \mathrm{trace}\Bigg(\mathbf{Q}\int_0^\infty\int_0^\infty \boldsymbol{\Phi}(\sigma_1)\mathbf{B}_d E[\mathbf{F}_t(t-\sigma_1) \\ &\quad \times \mathbf{F}_t^T(t-\sigma_2)]\mathbf{B}_d^T\boldsymbol{\Phi}^T(\sigma_2)\mathrm{d}\sigma_1\,\mathrm{d}\sigma_2\Bigg). \end{aligned} \tag{14}$$

Expectation $E[\mathbf{F}_t(t-\sigma_1)\mathbf{F}_t^T(t-\sigma_2)]$ contains the correlations in time between the disturbing forces. If these forces are mutually uncorrelated, the matrix is diagonal. If the signals are uncorrelated in time, the matrix is only nonzero if $\sigma_1 = \sigma_2$, in which case, the expectation is a constant matrix, $\mathbf{E}_{\mathbf{FF}^T}$, and Eq. (14) simplifies considerably,

$$J = \mathrm{trace}\left(\mathbf{Q} \int_0^\infty \mathbf{\Phi}(\sigma) \mathbf{B}_d \mathbf{E}_{\mathbf{FF}^{\mathrm{T}}} \mathbf{B}_d^T \mathbf{\Phi}^T(\sigma) \mathrm{d}\sigma \right) \tag{15}$$

or

$$J = \mathrm{trace}\left(\int_0^\infty \mathbf{\Phi}^T(\sigma) \mathbf{Q} \mathbf{\Phi}(\sigma) \mathrm{d}\sigma \mathbf{B}_d \mathbf{E}_{\mathbf{FF}^{\mathrm{T}}} \mathbf{B}_d^T \right). \tag{16}$$

No specific assumptions have so far been made on the location of the excitation. As can be seen from Eq. (3), the location would influence the matrix $\mathbf{B}_d$ and thus also the cost functions. In this analysis, the influence of the location of the excitation is an undesirable complication. If it is assumed that the excitation is a spatially completely random pressure field, this is equivalent to assuming that all the modes of the structure are excited equally, but in an uncorrelated fashion. This can be shown by analyzing the correlation between the excitation of the different modes.

Equation (4) describes how the modes are excited by point forces. Consider now the excitation of a single mode mn, $f_{mn}(t)$, by a pressure field, $p(x,y)$,

$$f_{mn}(t) = \frac{4}{M} \int_0^{l_y} \int_0^{l_x} \Psi_{mn}(x,y) p(x,y,t) \mathrm{d}x \, \mathrm{d}y. \tag{17}$$

The correlation between a mode kl and mode mn would then be, at any point in time

$$\begin{aligned} E[f_{kl} f_{mn}] &= \frac{16}{M^2} E\left[\int_0^{l_y} \int_0^{l_x} \Psi_{kl}(x_1,y_1) p(x_1,y_1,t) \mathrm{d}x_1 \, \mathrm{d}y_1 \right. \\ &\quad \left. \times \int_0^{l_y} \int_0^{l_x} \Psi_{mn}(x_2,y_2) p(x_2,y_2,t) \mathrm{d}x_2 \, \mathrm{d}y_2 \right] \\ &= \frac{16}{M^2} \int_0^{l_y} \int_0^{l_x} \int_0^{l_y} \int_0^{l_x} \Psi_{kl}(x_1,y_1) \Psi_{mn}(x_2,y_2) \\ &\quad \times E[p(x_1,y_1,t) p(x_2,y_2,t)] \mathrm{d}x_1 \, \mathrm{d}y_1 \, \mathrm{d}x_2 \, \mathrm{d}y_2. \end{aligned} \tag{18}$$

For a spatially completely random pressure field, there is no correlation between the pressures at two different locations and $E[p(x_1,y_1,t)p(x_2,y_2,t)]$ is equal to $\delta(x_1-x_2)\delta(y_1-y_2)E[p(x_1,y_1,t)^2]$. It is furthermore assumed that $E[p(x,y,t)^2]$ is constant for different (x,y). Equation (18) can then be rewritten as

$$\begin{aligned} E[f_{kl} f_{mn}] &= \frac{16}{M^2} \int_0^{l_y} \int_0^{l_x} \Psi_{kl}(x,y) \Psi_{mn}(x,y) E[p(x,y,t)^2] \mathrm{d}x \, \mathrm{d}y \\ &= \frac{16}{M^2} \int_0^{l_y} \int_0^{l_x} \Psi_{kl}(x,y) \Psi_{mn}(x,y) \mathrm{d}x \, \mathrm{d}y \, E[p(x,y,t)^2]. \end{aligned} \tag{19}$$

For the assumed, orthogonal mode shapes, the integral $\int_0^{l_y} \int_0^{l_x} \Psi_{kl}(x,y) \Psi_{mn}(x,y) \mathrm{d}x \, \mathrm{d}y$ is nonzero only if $kl=mn$, when it is equal to $l_x l_y/4$.

If the pressure field is assumed to have a white spectrum in time, as well as space, then Eqs. (19) and (16) can be combined,

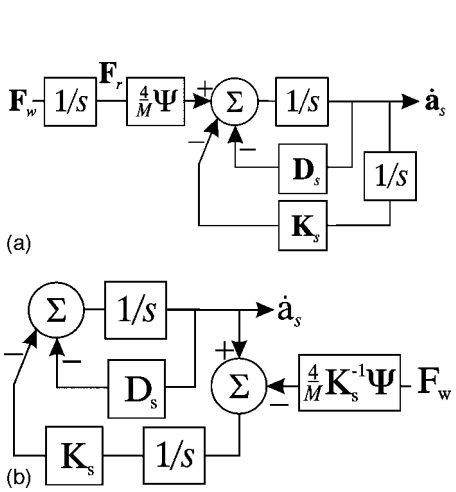


FIG. 2. Equivalent ways of modeling red noise, if velocity sensors and velocity based cost functions are used.

$$J = \mathrm{trace}\left(\int_0^\infty \mathbf{\Phi}^T(\sigma) \mathbf{Q} \mathbf{\Phi}(\sigma) \mathrm{d}\sigma \, \mathbf{P}_w \right) \tag{20}$$

with

$$\mathbf{P}_w = \left(\frac{4}{M} \right)^2 \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \tag{21}$$

Here, the magnitude of the expectation $E[p(x,y,t)^2]$ is chosen to be equal to $4/l_x l_y [N^2/m^4]$ and $\Phi(t)$ was assumed to only describe the structural states, with the state space vector $\mathbf{x}=[\mathbf{a}_s^T \dot{\mathbf{a}}_s^T]^T$. With this model, each of the structural modes is excited and also taken into account in the cost function. If acoustic radiation is examined, the matrix $\mathbf{P}_w$ needs to be extended with zeros for the states of the acoustic filters. The same should be done for the states of the controller if a dynamic velocity feedback controller is included in the states of the system. Spatially and temporaly uncorrelated excitation can be used as a model for a turbulent boundary layer excitation.

To model nonwhite noise, generally, extra states should be added to the system to color the noise before it acts on the system. However, for red noise, i.e., integrated white noise, in a system using velocity sensors and a velocity dependent cost function it can be shown that the shaping of the white noise can also be done without adding extra states, as can be seen from Figs. 2(a) and 2(b). The following matrix can then be used to calculate the red noise excitation,

$$\mathbf{P}_r = \left(\frac{4}{M} \right)^2 \begin{bmatrix} (\mathbf{K}_s^{-1})^T \mathbf{K}_s^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{22}$$

where $\mathbf{K}_s$ is part of the matrix in Eq. (4). This matrix should again be extended with zeros if the acoustic filters or a dynamic controller are taken into account. Note that, though the integrated white noise signal will tend to infinity at low frequencies, the response of the system in terms of modal velocity and radiated sound power, tends to 0. In Sec. V B 1 it is shown that the cost function in Eq. (20) can be evaluated using a solution of a Lyapunov equation as long as the system is asymptotically stable. In that case both cost functions remain bounded.

V. CONTROLLER DESIGN

In addition to the variations in the spectrum of the excitation and cost function, three different controller structures will be examined here; centralized constant gain velocity feedback control, decentralized constant gain velocity feedback control and LQG control. For all three control structures it is assumed that absolute velocity sensors are used with colocated point force actuators. All components are assumed ideal and without time delays.

A. Linear quadratic Gaussian control

An LQG controller [Kalman and Bucy (1961)] consists of two parts. One is the state estimator, which uses measured data to estimate the states of the system under control as accurately as possible. The other part is a feedback gain matrix, with the estimated states as input, and the control signal as outputs. The requirements for the use of LQG control, namely, stabilizability and detectability are met for this problem, as the system under control contains no unstable modes.

The state estimator gives an minimum variance estimation of the states, assuming a white noise excitation acts on the system and an uncorrelated white noise is added to the measured signals. If the excitation is not white noise, extra filters can be added to the model to shape the excitation. However, these filters should then also be included in the state estimation and increase the complexity of the controller.

Under the assumptions set out in Sec. IV for the white, spatially randomly distributed excitation, the matrix needed to calculate the effect of the excitation is the matrix $\mathbf{P}_w$. The weighting of the sensor noise is chosen to be small compared to the excitation and was set at 1×10^{-4} $(\mathrm{m}^2/\mathrm{s}^2)$ for each sensor for the white excitation. For red noise, the matrix $\mathbf{P}_r$ has to be used instead of $\mathbf{P}_w$. Also, the sensor signals are much smaller and to keep the sensor noise small in comparison to the measured signal, the weighting of the sensor noise is set equal to 1×10^{-10} $(\mathrm{m}^2/\mathrm{s}^2)$.

The other part of the controller, the feedback gain matrix, would minimize the cost function in Eq. (10) if full state information is available. Further details concerning the design of the LQG controller can be found, for example, in Skogestad and Postlethwaite (1996).

B. Constant gain controllers

When optimizing output feedback controllers, it is difficult to prove that a local minimum of the cost function is also the global minimum. This is discussed by Levine and Athans (1970), who examines constant gain, output feedback specifically. In this paper an algorithm is presented to find a minimum of the cost function, but it is also noted this algorithm is not guaranteed to converge. A different algorithm is used here, which is discussed by Anderson and Moore (1971) and is essentially a gradient descent algorithm. This section describes how this algorithm is implemented for both the centralized and the decentralized constant gain controllers.

1. Centralized constant gain controller

Levine and Athans (1970) examined a general state space model,

$$\dot{\mathbf{x}}(t)=\mathbf{A}\mathbf{x}(t)+\mathbf{B}_u\mathbf{u}(t),$$

$$\mathbf{y}(t)=\mathbf{C}\mathbf{x}(t) \tag{23}$$

with constant output feedback gain

$$\mathbf{u}(t)=-\mathbf{G}\mathbf{y}(t). \tag{24}$$

A cost function was used, which was similar to the one in Eq. (10). It was noted that, if the system starts at an initial set of states $\mathbf{x}(0)$ and no further external excitation acts on the system, the state at any particular time is given by

$$\mathbf{x}(t)=\mathbf{\Phi}(t)\mathbf{x}(0), \tag{25}$$

where $\mathbf{\Phi}(t)$ is the fundamental transition matrix for the controlled system, as defined by Eq. (13). For this constant gain output feedback controller, the matrix $\mathbf{A}_c$ is defined as

$$\mathbf{A}_c=\mathbf{A}-\mathbf{B}_u\mathbf{G}\mathbf{C}. \tag{26}$$

Combining Eqs. (25), (24), and (10) results in the equation

$$J=\mathbf{x}^T(0)\int_0^\infty \mathbf{\Phi}^T(t)[\mathbf{Q}+\mathbf{C}^T\mathbf{G}^T\mathbf{R}\mathbf{G}\mathbf{C}]\mathbf{\Phi}(t)\mathrm{d}t\,\mathbf{x}(0)$$

$$=\mathrm{trace}\left(\int_0^\infty \mathbf{\Phi}^T(t)[\mathbf{Q}+\mathbf{C}^T\mathbf{G}^T\mathbf{R}\mathbf{G}\mathbf{C}]\mathbf{\Phi}(t)\mathrm{d}t\,\mathbf{x}(0)\mathbf{x}^T(0)\right). \tag{27}$$

It can be easily seen that this is similar to Eq. (20), if $\mathbf{x}(0)\mathbf{x}^T(0)$ is replaced by $\mathbf{P}_w$. $\mathbf{P}_r$ is used when studying red noise excitation. If the system is asymptotically stable and the matrix $\mathbf{Q}$ is positive semidefinite, the cost function is bounded and equal to [Kalman and Bertram (1960)]

$$J=\mathrm{trace}(\mathbf{K}\mathbf{P}_w) \tag{28}$$

with $\mathbf{K}$ the positive definite solution of the Lyapunov equation

$$\mathbf{K}[\mathbf{A}-\mathbf{B}_u\mathbf{G}\mathbf{C}]+[\mathbf{A}-\mathbf{B}_u\mathbf{G}\mathbf{C}]^T\mathbf{K}+[\mathbf{Q}+\mathbf{C}^T\mathbf{G}^T\mathbf{R}\mathbf{G}\mathbf{C}]=\mathbf{0}. \tag{29}$$

The derivative of the cost function, J, with respect to the elements of the feedback gain matrix, $\mathbf{G}$, is equal to

$$\frac{\partial J}{\partial \mathbf{G}}=2\mathbf{R}\mathbf{G}\mathbf{C}\mathbf{L}\mathbf{C}^T=2\mathbf{B}_u^T\mathbf{K}\mathbf{L}\mathbf{C}^T, \tag{30}$$

where $\mathbf{K}$ is the solution of Eq. (29) and $\mathbf{L}$ is the solution of

$$[\mathbf{A}-\mathbf{B}_u\mathbf{G}\mathbf{C}]\mathbf{L}+\mathbf{L}[\mathbf{A}-\mathbf{B}_u\mathbf{G}\mathbf{C}]^T+\mathbf{P}_w=\mathbf{0}. \tag{31}$$

Using the derivative of the feedback gain, a simple algorithm can be formulated that will converge to a minimum on the cost function, if started at an initial stabilizing controller $\mathbf{G}_0$. The algorithm used here is similar to that in Anderson and Moore (1971):

(i) Calculate the cost J_k using Eqs. (28) and (29). If $k=0$, use the initial stabilizing controller $\mathbf{G}_0$.

(ii) Calculate the derivative of the cost function $\partial J_k/\partial \mathbf{G}_k$, using Eq. (30).

(iii) Update the gain matrix $\mathbf{G}_k$ according to

$$\mathbf{G}_{k+1} = \mathbf{G}_k - \frac{\epsilon}{\mathbb{F}[\partial J_k/\partial \mathbf{G}_k]} \frac{\partial J_k}{\partial \mathbf{G}_k}, \tag{32}$$

where ϵ is a small value to regulate the stepsize and $\mathbb{F}$ denotes the Frobenius norm. This norm is included to keep the stepsize in G_k independent of the size of the values in $\partial J_k/\partial \mathbf{G}_k$.

(iv) Check that the system is stable at these new gains and, if that is the case, calculate the cost J_{k+1}. If the system is no longer stable or $J_{k+1} > J_k$, reduce the stepsize ϵ, because the update has overshot the stability margins or an area where the cost is lower. Repeat the previous step and this step, until the system is stable and $J_{k+1} < J_k$, then repeat from the beginning.

(v) To stop the optimization, a suitable criterion can be chosen, such as a sufficiently small update in the gains, or a sufficiently small improvement in the cost function.

Though it cannot be proved that the algorithm converges to a global minimum, it is found in practice that the controller does converge to the same set of gains, independent of the choice of initial controller, the only exception is if the control locations are extremely close together [Engels and Elliott (2006)]. Therefore, for ease of formulation, applying the above algorithm will be referred to as *optimization*.

2. Decentralized constant gain controller

Geromel and Bernussou (1979) discussed the optimization of a constant gain decentralized controller, for the same system and cost functions as Levine and Athans (1970). The same algorithm can be used as for centralized control, except that the initial stabilizing control matrix should be diagonal and that the derivative with respect to the elements of the gain matrix is equal to

$$\frac{\partial J}{\partial \mathbf{G}} = \mathrm{diag}[2\mathbf{RGCLC}^T - 2\mathbf{B}_u^T\mathbf{KLC}^T], \tag{33}$$

where diag denotes a function that sets all off-diagonal terms of the matrix to 0.

VI. RESULTS

As noted above, it is important to compare the performance when equal amounts of control effort are used. In this study 16 equally spaced control locations are used, as indicated in Fig. 1. At each control location, ideal velocity sensors are assumed that are colocated with ideal force actuators.

Figures 3(a) and 3(b) show the resulting expectation of the kinetic energy and acoustic radiation when each of the three controllers are optimized for kinetic energy and acoustic radiation, respectively. A white noise excitation is assumed and the control effort weighting was adjusted such that the expected controller effort was equal to 300 N^2 for each controller. The kinetic energy and sound power density around the first resonance frequency is reduced by about 25 dB dropping off to about 10 dB reduction at other resonances. Despite the seemingly better performance of LQG control in Fig. 3(a), the integral over time [Eq. (20)] of kinetic energy was reduced by 5.5 dB for *each* of the controllers. For sound power, the reductions were 4.5 dB for LQG control, 4.1 dB for centralized constant gain control, and 4.0 dB for decentralised constant gain control.

Figures 3(c) and 3(d) also show kinetic energy and acoustic radiation, but with a red noise excitation instead of white noise. The level of excitation is far less in this case and so the control effort was now limited to 3×10^{-3} N^2 for each controller. The reductions in kinetic energy for the different controllers at this effort are 22.3 dB for LQG control and 23.3 dB for both centralized and decentralized constant feedback gain control. For sound power the reductions are, respectively, 19.7, 22.6, and 22.3 dB.

In this case LQG control performs less well than the constant gain controllers. This is due to the extent of the reductions that are achieved, which cause the sensor signals to be in the range of the sensor noise that was assumed in the design of the LQG controller. Though the level of the reductions achieved with this control effort are not obtainable in a real situation due to the presence of sensor noise, the level of control effort was chosen to emphasize the difference between the constant gain and LQG controllers. The reductions that are achieved are much larger than in the white noise case because most of the cost in the cost function is caused by the energy in the first mode.

It is interesting to note that each of the gains in the optimized constant gain decentralized controller were of similar magnitude, but that the optimized centralized constant gain controller also had off-diagonal gains that were of the same order of magnitude as the on-diagonal gains. Dispite the magnitude of these gains, they did not contribute significantly to the reduction in the cost function.

The overall difference in the cost function is difficult to see from these plots and it is not clear whether this level of effort is most appropriate. Therefore, the overall reduction in the expected kinetic energy and sound radiation reduction should be examined as a function of control effort. These values can be evaluated using Eqs. (28) and (29). Figure 4(a) and 4(b) show the change in kinetic energy and radiated sound power as a function of the control effort for white noise excitation, which has been computed by optimizing the three controllers with varying control effort weightings. Still higher control efforts can be achieved by using higher gains in the control loops, but this results in worse, rather than better performance. It can be seen that there is some advantage in using LQG control rather than constant gain feedback, since for a given performance, the control effort is slightly lower, but this difference is small. It should be noted that for the control of sound power, the LQG controller requires a total of 347 states.

Figures 4(c) and 4(d) show the results when the system is excited by red noise. There are nearly no differences between the controllers for low control efforts, but the LQG controller performs worse at higher control efforts, which is again due to the fact that the LQG controller takes sensor

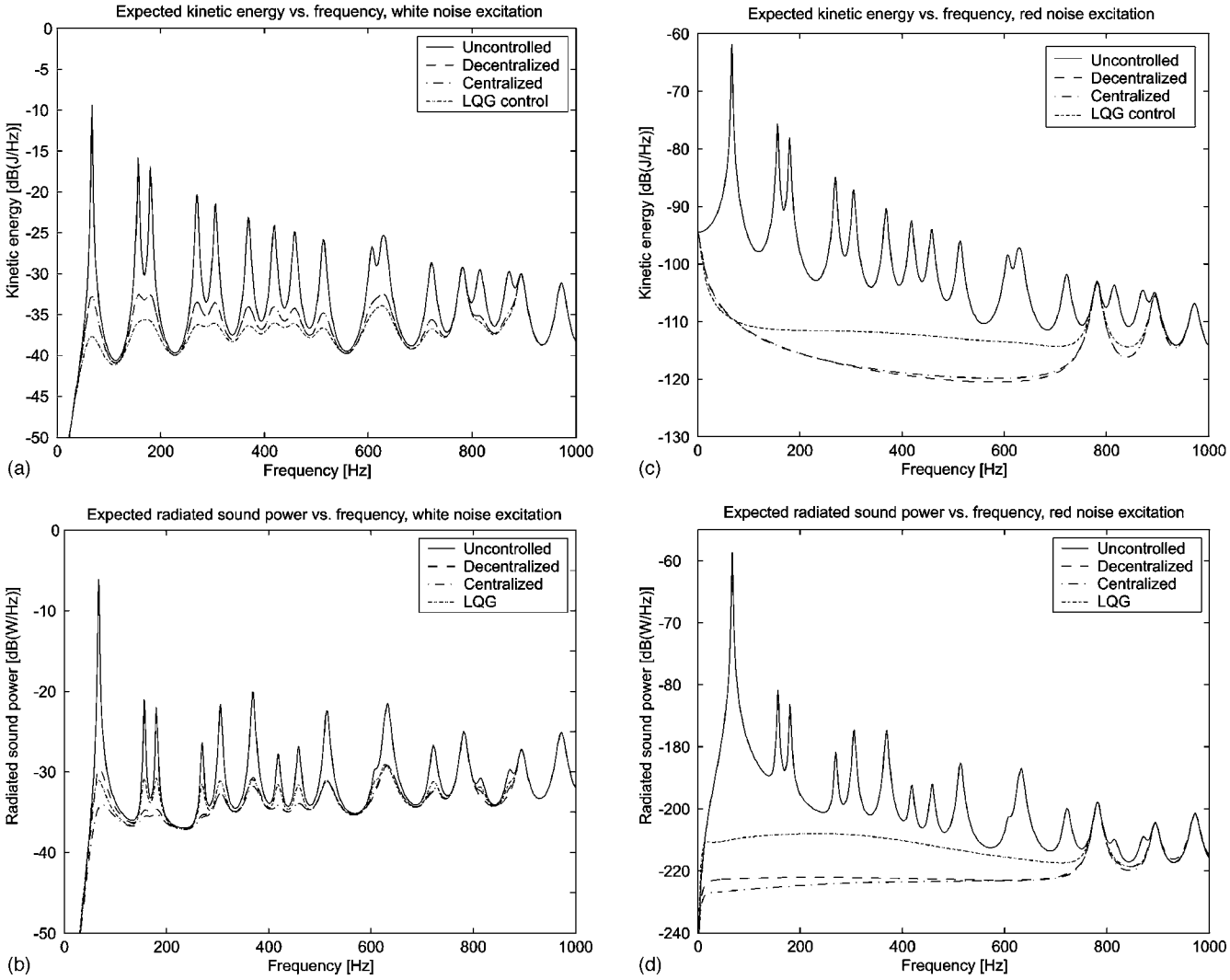


FIG. 3. Spectrum of expected kinetic energy (KE) and radiated sound power (ac), assuming white or red noise excitation, before and after control, using different controllers optimized to control kinetic energy and radiated sound power, respectively. The expected average control effort for each controller was limited to 300 N^2 for white noise excitation and 3×10^{-3} N^2 for red noise. (a) KE vs frequency, white noise. (b) ac vs frequency, white noise. (c) KE vs frequency, red noise. (d) ac vs frequency, red noise.

noise into account in the design, but that no noise was taken into account in the calculation of the cost function. These graphs could also provide an important design tool in determining the correct trade off between performance and control effort for a given application.

From these figures it can be seen that there is little difference between the effectiveness of the different controllers. For a white noise excitation there is some advantage in using LQG control over constant gain control if acoustic radiation is considered. However, these results were obtained with the controller at particular points on the plate, where certain modes cannot be controlled, as can be seen in Fig. 3(a). It is not clear how this affects the results obtained and whether the same conclusions can be drawn for a different placement of the actuators. Therefore the difference between the controllers for a given control effort has also been examined for randomly placed control locations. The number of control locations was limited to 5. It was found that the differences between centralized, decentralized constant gain and LQG control remain small, though LQG control did show slightly better performance in the case of white noise for both kinetic energy and acoustic radiation. LQG control did not improve the performance in the case of red noise excitation, because of the amount of sensor noise considered in the design of the controller.

VII. CONCLUSIONS

For the model problem considered, with 16 colocated velocity sensors and point force actuators, there is little performance gain in using centralized static feedback gain control or LQG control over decentralized static feedback gain control, when similar amounts of control effort are used. LQG control gives some improvement in performance in the case of a white, randomly distributed excitation, but may not be worth the added complexity of the controller. Similar results were also observed for five channel controllers with randomly located control locations.

This seems to contrast with the work of Smith and Clark (1998) who saw a significant improvement at some frequen-

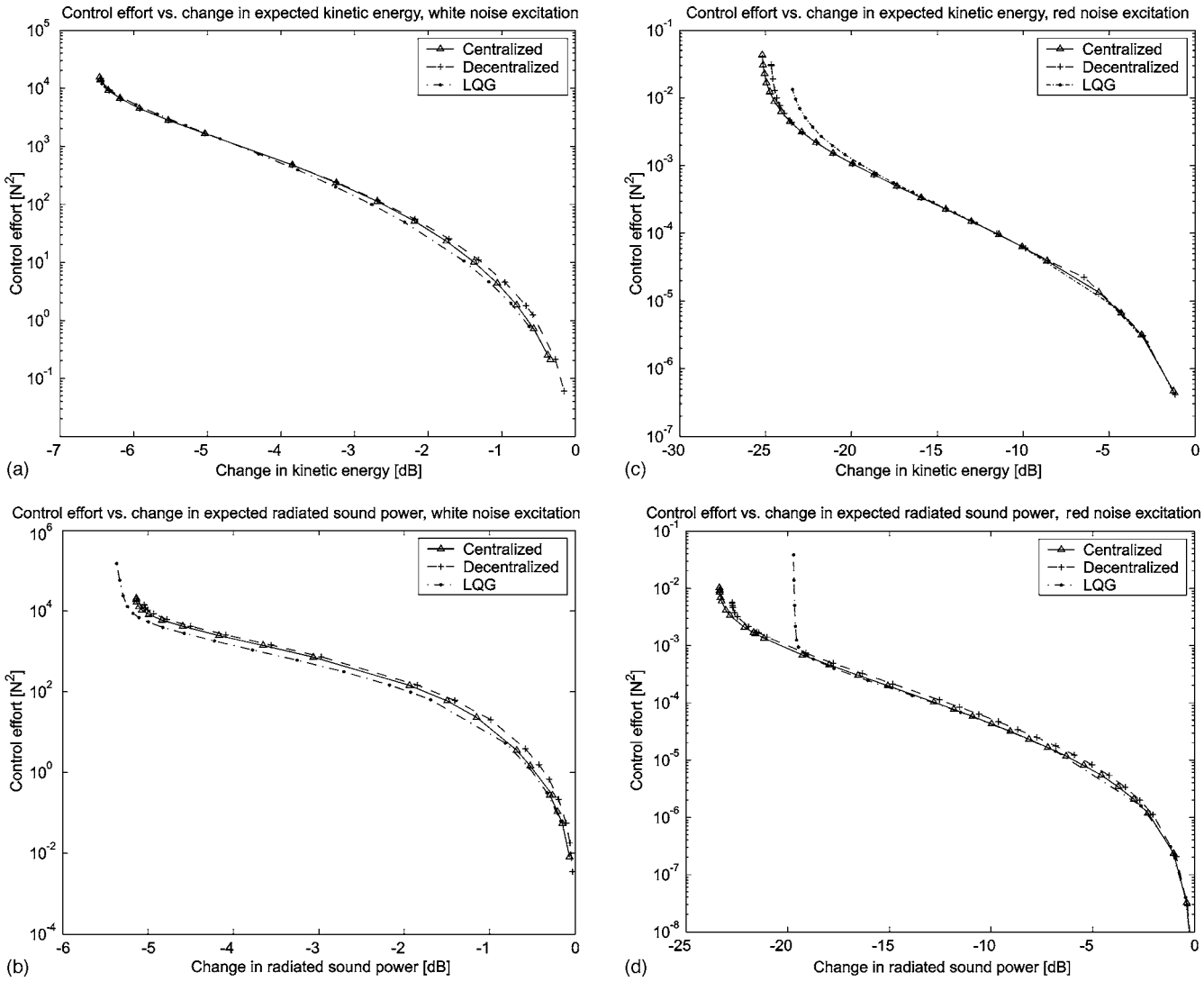


FIG. 4. Change in expected kinetic energy (KE) and radiated sound power (ac) as a function of effort, assuming white or red noise excitation, using different controllers optimized to control kinetic energy and radiated sound power, respectively. (a) KE vs effort, white noise. (b) ac vs effort, white noise. (c) KE vs effort, red noise. (d) ac vs effort, red noise.

cies, when using LQG control in a single channel controller and a large distributed sensoriactuator. The frequency range over which this improvement occurred was limited though and results, on average, in only a small difference. Elliott (2004) has noted that if the number of control loops is equal to the number of controlled modes, then under idealized circumstances the effect of a decentralized constant gain controller would be equal to that of a fully coupled modal controller. Although in these simulations, 60 structural modes were taken into account in both the model and the cost function, the total kinetic energy is dominated by a significantly smaller number of modes, as can be seen from Fig. 3(a) for white noise excitation, and an even smaller number for red noise excitation, Fig. 3(c). This may explain why, in the case of red noise excitation, when there are only very few modes contributing significantly to the cost function, there is hardly any difference between centralized, decentralized, and LQG control. Further research into the relation between the number of excited modes and the number of control locations for which the difference in performance between dynamic controllers and static controllers becomes small, may prove interesting. Using piezoceramic actuators can be more practical than a point force actuator and it would be interesting to see if the use decentralized, constant gain control is as effective in that case as here. The stability of constant gain feedback controllers using such actuators does require careful attention [Gardonio and Elliott (2004)].

ACKNOWLEDGMENTS

This work has been partly funded by the European Community InMAR project and partly by the Engineering and Physical Sciences Research Council (EPSRC). The work of one of the authors (O.B.) is sponsored by the Data Information Fusion Defence Technology Centre. Furthermore, the authors would like to thank Michel Verhaegen for useful discussions on this topic.

Anderson, B. D. O., and Moore, J. B. (**1971**). *Linear Optimal Control* (Prentice-Hall, Englewood Cliffs, NJ).

Balas, M. J. (**1979**). "Direct velocity feedback control of large space structures," J. Guid. Control **2**, 252–253.

Baumann, W. T., Saunders, W. R., and Robertshaw, H. H. (**1991**). "Active

suppression of acoustic radiation from impulsively excited structures," J. Acoust. Soc. Am. **90**, 3202–3206.
Bingham, B., Atalla, M. J., and Hagood, N. W. (**2001**). "Comparison of structural-acoustic designs on an active composite panel," J. Sound Vib. **244**, 761–778.
Borgiotti, G. V. (**1990**). "The power radiated by a vibrating body in an acoustic fluid and its determination from boundary measurements," J. Acoust. Soc. Am. **88**, 1884–1893.
Borgiotti, G. V., and Jones, K. E. (**1994**). "Frequency independence property of radiation spatial filters," J. Acoust. Soc. Am. **96**, 3516–3524.
Clark, R. L., and Cox, D. E. (**1997**). "Multi-variable structural acoustic control with static compensation," J. Acoust. Soc. Am. **102**, 2747–2756.
Clark, R. L., and Frampton, K. D. (**1999**). "Aeroelastic structural acoustic control," J. Acoust. Soc. Am. **102**, 743–754.
Cox, D. E., Gibbs, G. P., Clark, R. L., and Vipperman, J. S. (**1998**). "Experimental robust control of structural acoustic radiation," *39th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference* (AIAA).
Elliott, S. J. (**2004**). "Distributed control of sound and vibration," Proceedings of ACTIVE04.
Elliott, S. J., Gardonio, P., Sors, T. C., and Brennan, M. J. (**2002**). "Active vibroacoustic control with multiple local feedback loops," J. Acoust. Soc. Am. **111**, 908–915.
Elliott, S. J., and Johnson, M. E. (**1993**). "Radiation modes and the active control of sound power," J. Acoust. Soc. Am. **94**, 2194–2204.
Engels, W. P., Baumann, O. N., and Elliott, S. J. (**2004**). "Centralized and decentralized feedback control of kinetic energy," Proceedings of ACTIVE04.
Engels, W. P., and Elliott, S. J. (2006). Optimal velocity feedback control on a beam (unpublished).
Fuller, C. R., Elliott, S. J., and Nelson, P. A. (**1996**). *Active Control of Vibration* (Academic, London).
Fuller, C. R., Kidner, M., Li, X., and Hansen, C. H. (**2004**). "Active-passive heterogeneous blankets for control of vibration and sound radiation," Proceedings of ACTIVE04.
Gardonio, P., Bianchi, E., and Elliott, S. (**2004**). "Smart panel with multiple decentralized units for the control of sound transmission. Part I: theoretical predictions," J. Sound Vib. **274**, 163–192.
Gardonio, P., and Elliott, S. (**2004**). "Smart panels for active structural acoustic control," Smart Mater. Struct. **13**, 1314–1336.
Geromel, J. C., and Bernussou, J. (**1979**). "An algorithm for optimal decentralized regulation of linear quadratic interconnected systems," Automatica **15**, 489–491.
Gibbs, G. P., Clark, R. L., Cox, D. E., and Vipperman, J. S. (**2000**). "Radiation modal expansion: Application to active structural acoustic control," J. Acoust. Soc. Am. **107**, 332–339.
Kalman, R. E., and Bertram, J. E. (**1960**). "Control system analysis and design via the "second method" of Lyapunov, I Continuous-time systems," J. Basic Eng. **82**, 371–393.
Kalman, R. E., and Bucy, R. S. (**1961**). "New results in linear filtering and prediction theory," J. Basic Eng. **83**, 95–108.
Levine, W. S., and Athans, M. (**1970**). "On the determination of the optimal constant output feedback gains for linear multivariable systems," IEEE Trans. Autom. Control **AC-15**, 44–48.
Maury, C., Gardonio, P., and Elliott, S. J. (**2002**). "A wavenumber approach to modeling the response of a randomly excited panel, part I: general theory," J. Sound Vib. **252**, 83–113.
Meirovitch, L. (**1986**). *Elements of Vibration Analysis*, 2nd ed. (McGraw-Hill, New York).
Skogestad, S., and Postlethwaite, I. (**1996**). *Multivariable Feedback Control—Analysis and Design* (Wiley, Chichester, England).
Smith, G. C., and Clark, R. L. (**1998**). "The influence of frequency-shaped cost functionals on the structural acoustic control performance of static, output feedback controllers," J. Acoust. Soc. Am. **104**, 2236–2244.
Sun, J. Q. (**1996**). "Some observations on physical duality and colocation of structural control sensors and actuators," J. Sound Vib. **194**, 765–770.
Thomas, D. R., and Nelson, P. A. (**1995**). "Feedback control of sound radiation from a plate excited by a turbulent boundary layer," J. Acoust. Soc. Am. **98**, 2561–2662.
Vipperman, J. S., and Clark, R. (**1991**). "Implications of using colocated strain-based transducers for output active structural acoustic control," J. Acoust. Soc. Am. **106**, 1392–1399.

Stability analysis of active control of self-sustained pressure fluctuations due to flow over a cavity

P. Micheau,[a] Ludovic Chatellier, Janick Laumonier, and Yves Gervais
Laboratoire d'Etudes Aérodynamiques, Bat. K, 40 av Recteur Pineau, 86000 Poitiers, France

(Received 29 June 2005; revised 5 December 2005; accepted 6 December 2005)

The problem addressed in this paper is the stability analysis of self-sustained pressure fluctuations due to low Mach number turbulent flow over a large shallow cavity in a wing-shaped model. In the studied configuration, sound pressure was generated by the instability of the mixing layer. An original active device, a vibrating surface located at the trailing edge, and an adjustable narrow-band controller were used to control the self-oscillation: to amplify it in an open loop, and to attenuate it in a closed loop. A complete analysis of the system based on the describing function analysis is presented. Experimental data, obtained in an Eiffel-type wind-tunnel compare favorably with the prediction of the model. With the active control, the instability is controlled: the pressure fluctuations measured in the cavity are attenuated to the ground level. Moreover, with the proposed model and the original device, the equivalent volume velocity of the unstable mixing layer and the acoustic feedback gain can be experimentally estimated.
 [DOI: 10.1121/1.2161635]

PACS number(s): 43.50.Ki, 43.28.Ra [TDM] Pages: 1496–1503

I. INTRODUCTION

The problem addressed in this paper is to control self-sustained pressure fluctuations due to low Mach number flow over an open cavity. A flow over open cavities can induce self-sustained hydrodynamic oscillations of the turbulent mixing layer that produces pressure fluctuations. When the frequency of the hydrodynamic oscillation is close to an acoustic mode, an intense coupling occurs between the flow oscillation and the pressure fluctuation inside the cavity that leads to a high level narrow-band noise. Such self-oscillation can be desirable, with musical instruments for example, or unwanted, as with the open sunroof of a moving car.

Many passive solutions have been proposed to solve the problem of self-oscillating pressure: leading-edge spoilers, trailing-edge deflectors, and leading-edge flow diffusers.[1] More recently, active flow control has been proposed,[2] following the results obtained in active sound control when a loudspeaker is used to generate antinoise in order to attenuate an unwanted noise.[3] Many devices have been used in active control of flow: upstream air injection method, oscillating spoilers, loudspeakers, or active flaps.[4–6] The originality of the active device presented in this paper (Fig. 1) is the use of a volume velocity source located along the trailing edge to actively control the self-oscillating pressure induced by the flow.[7] The main contribution of this paper is to interpret the action of the active device on the self-oscillating phenomena.

The vortex sound theory describes how sound can be induced by periodic shedding of discrete vortices at the leading edge.[8] The interaction of vortices with the downstream edge, after their convection across the cavity, generates an acoustic pressure fluctuation that triggers the vortex shedding upstream of the leading edge. This closed loop leads to the reinforcement of the frequency and then the generation of a self-sustained fluctuation. However, it may be argued that a vortex-edge interaction occurring at a low Mach number is not likely to produce an intense upstream pressure perturbation, and that self-sustained oscillations can be a direct consequence of the natural instability of the mixing layer, intensified by the presence of an obstacle.[9] This theory was validated with the experimental setup presented in this paper.[10] However, this theory predicts the growth of instabilities and must invoke nonlinear factors to limit them.

In order to model self-oscillation due to a nonlinear feedback loop, the describing function theory can be used.[11] Previous works have shown that this theory can be applied to predict the frequency and the relative amplitude of the cavity pressure fluctuations for a range of flow velocities.[12,13] Because this approach is developed within the framework of the feedback loop analysis, it is well adapted to interpret the action of the active device on the acoustic feedback.

For this purpose, the paper is organized as follows. In Sec. II we present the proposed model of the self-oscillator and its analysis with the describing function theory. In Sec. III we introduce the active device, its controller, and its analysis. In Sec. IV we show the experimental results obtained in a wind-tunnel environment.

II. MODEL OF THE SELF-OSCILLATION

A. Strouhal number prediction

The vortex sound theory describes the feedback mechanism in cavities in grazing flows with the shedding of vortices from the upstream edge. Due to the free-stream velocity, U_∞, the vortices are convected over the cavity length, L, at the speed $U_c = \kappa U_\infty$ where κ is typically from 0.3 to 0.6. When a vortex reaches the downstream edge, it generates a

[a]Corresponding author: G.A.U.S., Department of Mechanical Engineering, Université de Sherbrooke, Sherbrooke, Québec, J1K 2R1, Canada. Phone: (819) 821-800, ext 2161; fax: (819) 821-7163; electronic mail: Philippe.Micheau@USherbrookre.ca

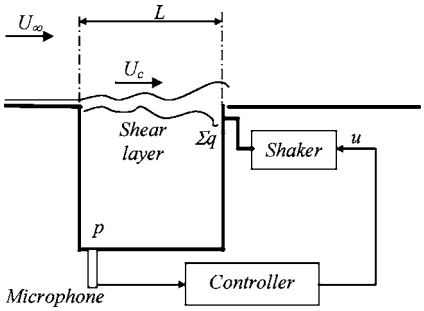


FIG. 1. Sketch of the cavity with the actuator generating an antivolume velocity at the trailing edge (p: pressure inside the cavity, u: command, U_∞: free-stream velocity, U_C: speed of the convected unstable shear layer, Σq: volume velocity of sources located at the trailing edge).

pressure pulse that propagates to the upstream edge at the sound speed, c, and initiates the shedding of a new vortex. The appropriate period of this cycle is then the sum of the vortex convection time, L/U_c, and the pressure pulse propagation time, L/c, divided by m, the number of vortices along the cavity length. Hence, the Strouhal number is given by the Rossiter relation:

$$S_m = \frac{f_m L}{U_\infty} = \frac{m-\alpha}{M+1/\kappa}, \tag{1}$$

with f_m the frequency of the self-oscillation, $M=U_\infty/c$ the Mach number, and $\alpha \approx 0.2$. Once the fitting parameters, κ and α, have been adjusted, most of the experimental results collapse well with this formula for subsonic flows (a corrected formula can also be applied to high subsonic and supersonic flows). However, some results gradually deviate from this model for Mach numbers lower than 0.3.

B. Describing function analysis

The describing function analysis is well adapted to predict the existence of the self-oscillation due to a nonlinear complex feedback loop. For this purpose, the feedback relation must be described by a single nonlinear element, $N(A)$, in series with a single linear element, $G(\omega)$. Figure 2 shows such a block diagram model. The fundamental assumption of the describing function analysis requires that only the fundamental component of each signal has to be considered. Hence, the form of the output signal is considered sinusoidal: $-p(t)=A\sin(\omega t+\Phi)$. As this sinusoidal output feeds into the nonlinear part (due to the feedback loop), the output of the nonlinear element, $q(t)$, is often periodic, but it is generally a nonsinusoidal function. However, only the fundamental component of $q(t)$ is considered, and its amplitude is written as $N(A)A$, where $N(A)$ is the describing function, an extension of the notion of frequency response.

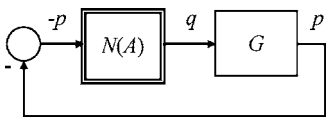


FIG. 2. Nonlinear feedback loop model with $N(A)$ the describing function of the nonlinear element and $G(j\omega)$ the transfer function of the linear element.

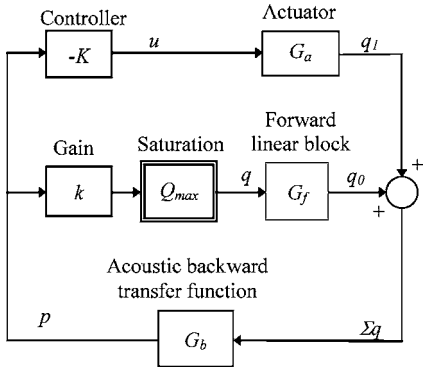


FIG. 3. Block diagram illustrating the model composed of the forward blocks (gain, saturation, and linear transfer function), the acoustic backward block, and the active control device (controller and actuator).

For a typical saturation nonlinearity of linearity range $A_{\max}$ and gain k, its describing function is

$$N(A) = \begin{cases} k, & \text{for } A < A_{\max}, \\ \dfrac{2k}{\pi}\left(\sin^{-1}\left(\dfrac{A_{\max}}{A}\right) + \dfrac{A_{\max}}{A}\sqrt{1-\dfrac{A_{\max}^2}{A^2}}\right) & \text{for } A \geqslant A_{\max}. \end{cases} \tag{2}$$

When the input amplitude is in the linearity range, $A < A_{\max}$, the describing function (2) is equal to a linear feedback of gain $N(A)=k$. In this case, the Nyquist criterion states that the closed loop system is stable if, and only if, the $-1/k+j0$ point is not encircled by the Nyquist curve of $G(\omega)$. On the other hand, the closed loop system is unstable if $G(\omega)$ encircles the $-1/k+j0$ point: a self-oscillating signal of frequency ω_m grows exponentially. However, due to the saturation, the exponential growth of the unstable oscillation is limited to amplitude $A_{\max}$. According to the describing function analysis, the self-sustained oscillation of amplitude A_m and frequency ω_m can exist if the following equality is reached:[10]

$$G(\omega_m) = -\frac{1}{N(A_m)}. \tag{3}$$

Equation (3) represents the intersection between $G(\omega)$ and $-1/N(A)$ in the complex plane at the frequency ω_m and for the amplitude A_m. Hence, graphical representations can be used to perform analysis and design.

C. Model of the aeroacoustic system

Figure 3 shows the block diagram of the aeroacoustic model, including the active device. The unstable shear layer is considered as an aeroacoustic source of volume velocity, q_0. The backward block, $G_b(\omega)$, represents the acoustic feedback that synchronizes the phenomenon. The transfer mechanism between the pressure inside the cavity, p, and the volume velocity generated by the unstable shear layer, q, is represented by a linear gain block, k, and a nonlinear saturation block. For small amplitude of the pressure, the relation is linear. On the other hand, for large amplitude of the pressure, the output of the nonlinearity is saturated: the natural dissipative mechanism prevents unlimited growth of the self-oscillation. The time delay due to the convection is repre-

sented by the forward block, $G_f(\omega)$. The actuator is an anti-aeroacoustic source of volume velocity, q_1 (a baffled vibrating surface), commanded by the controller.

D. The nonlinear forward blocks

The volume velocity, q_0, is driven by the pressure, p, via the nonlinear forward block and the linear forward block. For a frequency ω of self-oscillation, it can be written as

$$q_0(\omega) = G_f(\omega)N(A)p(\omega), \tag{4}$$

with $A=|p(\omega)|$ the pressure amplitude and $N(A)$ given by Eq. (2). The linear forward block, $G_f(\omega)$, represents the shear layer convection at the speed U_c across the cavity from the upstream edge to the trailing edge (over the distance L), and a phase shift of $\alpha 2\pi$ radians:

$$G_f(\omega) = \exp\left(-j\frac{\omega L}{U_c} - j\alpha 2\pi\right). \tag{5}$$

The model presented with Eqs. (4) and (5) means that the volume velocity amplitude is a constant value when the nonlinear block is saturated: $|q_0(\omega)| \approx 4kA_{\max}/\pi$ for $A \gg A_{\max}$. It also means that the volume velocity phase is synchronized with the pressure phase: $\angle q_0(\omega_m) = \angle p(\omega_m) + \Phi$ with $\Phi = -\omega L/U_c - \alpha 2\pi$.

E. The acoustic backward transfer function

The acoustic backward transfer function $G_b(\omega)$ is between the volume velocity source located at the trailing edge and the pressure fluctuations that trigger the instability. In the case of grazing flows over cavities, the Rossiter feedback mechanism assumes that the acoustic backward transfer function is modeled as the sound travel time at the sound speed from the downstream to the upstream edge of an open cavity.[8] On the other hand, for grazing flows over a Helmholtz resonator, the acoustic backward transfer function is approximated by a second-order system, because the pressure inside the cavity is related to the volume velocity source by a one-dimensional lumped-element oscillator.[11] However, in practice, the environment of the cavity can modify the acoustic backward transfer function.[5] In this paper, the proposed approach is to consider an experimental measurement of $G_b(\omega)$ in order to include both the cavity response and its environment. The basic assumption of this approach is to assume that the proposed active device, located at the downstream edge, is very close to the aeroacoustic source, $q_0(\omega)$. Consequently, the acoustic backward transfer function is defined between the volume velocity of the active device, $q_1(\omega)$, and the pressure fluctuation measured by the microphone, $p_1(\omega)$:

$$G_b(\omega) = \frac{p_1(\omega)}{q_1(\omega)}, \quad \text{when } q_0(\omega) = 0. \tag{6}$$

F. Strouhal number prediction

The Strouhal number prediction, Eq. (1), can be deduced from the reduced model. The acoustic backward transfer function, Eq. (6), in series with the forward transfer function, Eq. (5), gives the linear part:

$$G(\omega) = -|G_b(\omega)|\exp\left(-j\frac{\omega L}{U_c} + j\angle G_b(\omega) - j\alpha 2\pi\right). \tag{7}$$

According to Eq. (3), the existence of a self-oscillation (of frequency ω_m) requires that the linear part given by Eq. (7) have a null phase, $-\omega_m(L/U_c) + \angle G_b(\omega_m) - \alpha 2\pi = 0$ modulus 2π, which means a perfect reinforcement of the self-oscillation by the acoustic feedback (also called a positive feedback). With the Strouhal number, Eq. (1), the phase condition in Eq. (7) corresponds to the following Strouhal number prediction: $(2\pi/\kappa)S_m = \angle G_b(\omega_m) - 2\pi\alpha \bmod 2\pi$. By indexing m each solution from the smallest frequency ($0 < \omega_0 < \omega_1 < \cdots$), the phase condition corresponds to the following condition:

$$S_m = \kappa\left(m - \alpha + \frac{\angle G_b(\omega_m)}{2\pi}\right). \tag{8}$$

When the acoustic backward transfer function represents a time delay, $G_b(\omega) = \exp(-j\omega L/c)$, due to the acoustic propagation at the sound speed c from the trailing edge to the upstream edge (distance L), Eq. (8) leads to Eq. (1) because $\angle G_b(\omega_m)/2\pi = -MS_m$. Hence, the Rossiter feedback mechanism is a special case of the reduced-order model proposed in this paper. When the parameters κ and α are known, Eq. (8) can be used to predict the frequencies of possible self-oscillation for a backward acoustic transfer function. On the other hand, the parameters κ and α can be estimated with Eq. (8) from experimental measurements of ω_m vs U_∞.

III. MODEL OF THE ACTIVE CONTROL

A. Model of the active device

Figure 1 shows the original device used to control the self-oscillating pressure fluctuation: a vibrating surface located at the trailing edge. The main assumption of this paper is to consider that the volume velocity of the aeroacoustic source, q_0, has the same value with and without the action of the device (when the self-oscillation is still observable). In other words, the unsteady flow phenomenon is not *directly* perturbed by the action of the vibrating surface: the active device acts on the acoustic feedback mechanism, but not directly on the shear layer. With this assumption, the action of the active device can be modeled as a linear summation of the aeroacoustic volume velocity, q_0, with the antiaeroacoustic one, q_1:

$$\Sigma_q(\omega) = q_0(\omega) + q_1(\omega). \tag{9}$$

The vibrating surface is assumed to vibrate like a baffled rigid piston of area S; its volume velocity is given by $q_1(\omega) = (S/j\omega)\ddot{x}(\omega)$ where the surface acceleration, $\ddot{x}$, is driven by the shaker command, u. Hence, it can be defined a

linear transfer function, $G_a(\omega)$, between u, and the antiaeroacoustic volume velocity:

$$q_1(\omega) = G_a(\omega)u(\omega). \tag{10}$$

Consequently, Eqs. (9) and (10) represent a model of the active device. In order to complete the active control model, the equations of the controller must be included, i.e., the relation between the measured pressure fluctuation and the shaker command.

B. Open loop control

In open loop control, the method consists in commanding the shaker with a sinusoidal waveform at the forced frequency ω_0. In self-oscillating system, when the frequency of the open loop command, ω_0, is close to the frequency of the self-oscillation, ω_m, the self-oscillation may be synchronized to the command: the self-oscillation is forced at the frequency of the open loop command ($\omega_0=\omega_m$). The describing function theory says that the forced oscillation must be a solution of

$$\frac{p(\omega_0)}{p_1(\omega_0)} = \frac{1}{1+N(A)G(\omega_0)}, \tag{11}$$

with $A=|p(\omega_0)|$, $p(\omega_0)=p_1(\omega_0)+p_0(\omega_0)$, and $p_1(\omega_0)=G_b(\omega_0)q_1(\omega_0)$ the pressure generated by the active device. For a forced frequency, $\omega_0=\omega_m$, equal to the self-oscillated frequency, $\omega_0=\omega_m$, the right term of Eq. (11) is a real number because $G(\omega_m)=-|G(\omega_m)|$ due to Eq. (3). By considering that $|p_1(\omega_0)|=|G(\omega_0)||q_1(\omega_0)|$ because $|G(\omega)|=|G_b(\omega)|$, $\forall\omega$ due to Eq. (7), Eq. (11) becomes $A/|q_1(\omega_0)|=|G(\omega_0)|/[1-N(A)|G(\omega_0)|]$. Hence, it may be possible to identify the describing function with $N(A)=[1/|G(\omega_0)|]-[|q_1(\omega_0)|/A]$ and measurements of A vs $|q_1(\omega_0)|$. Furthermore, if we assume the special case of a saturated source, the describing function, Eq. (2), can be approximated by $N(A)\approx Q_{\max}/A$, where $Q_{\max}=4kA_{\max}/\pi$; hence, Eq. (11) leads to the following equation:

$$A = |G(\omega_0)|[|q_1(\omega_0)| + Q_{\max}]. \tag{12}$$

Consequently, when the aeroacoustic source is saturated in open loop control, the amplification of A vs $|q_1(\omega_0)|$ must respect Eq. (12).

C. Closed loop control

In closed loop control, the actuator command is synchronized with the pressure measured at the bottom of the cavity. In this paper, a bandpass filtering centered on the tone to be controlled is applied to the pressure signal feeds to the feedback controller. Consequently, the spillover effects due to the feedback loop are restrained in a narrow frequency band, thus avoiding any excitation from other tones. To apply this bandpass control strategy, a complex envelope controller was implemented; such an approach was used to realize controllers rejecting narrow band signals[14] and was successfully experimented in active control of pulsed flow[15] and in active vibration control.[16] With this approach, the controller is described by three equations:

$$P(t) = 2f(t) \otimes [p(t)\exp(-j\omega_0 t)], \tag{13}$$

$$u(t) = \mathrm{Re}[U(t)\exp(+j\omega_0 t)], \tag{14}$$

$$U(t) = -KP(t), \tag{15}$$

where $\mathrm{Re}(\cdot)$ denotes the real part of $(\cdot)$, $\otimes$ denotes the convolution operator, $f(t)$ is the impulse response of a low-pass filter, and $K=|K|\exp(j\psi)$ is a complex feedback gain. By considering the Fourier transforms of Eqs. (13)–(15), it can be demonstrated that the proposed controller is equivalent to a narrow band compensator centered on the central frequency ω_0:

$$\frac{q_1(\omega)}{p(\omega)} = -KG_a(\omega)f(\omega-\omega_0) - K^*G_a(-\omega)f(-\omega-\omega_0), \tag{16}$$

where K^* is the conjugate of K. Equation (16) shows that the low-pass filter, $f(\omega)$, can be used to select the controller frequency band. For a frequency ω very close to ω_0, Eq. (16) can be approximated by a complex gain because $f(\omega-\omega_0)\approx 1$ and $f(-\omega-\omega_0)\approx 0$:

$$\frac{q_1(\omega_0)}{p(\omega_0)} = -KG_a(\omega_0) = -K_a. \tag{17}$$

Equation (17) shows that the proposed controller is a useful and simple way to tune (with K) the phase shift and the magnitude of a narrow band compensator, Eq. (16), at the self-oscillation frequency, i.e., the complex gain K_a.

D. Analysis of the active control in closed loop

When the active device is commanded in closed loop control, the whole system (the flow and its active control) is equivalent to an autonomous nonlinear system. The analysis consists in determining the existence of self-oscillations and their characteristics: amplitude and frequency.

When a small feedback gain is used such that the volume velocity amplitude is not saturated, the controller compensates, or accentuates, the acoustic feedback. To demonstrate this effect, let us only consider small pressure perturbations (there is no physical saturation). Hence, Eqs. (2), (9), and (17) lead to the following equation:

$$\Sigma q_0(\omega) = [G_f(\omega)k - K_a]p(\omega). \tag{18}$$

When the phase shift of the feedback gain K_a is tuned to compensate the two phase lags in order to obtain a perfect negative feedback, $\angle K_a = \angle G_f(\omega)$, the two gains can be subtracted and Eq. (18) leads to $|\Sigma q(\omega)| = (k-|K_a|)A$. Hence, the critical point of the Nyquist criterion becomes $-1/(k-|K_a|)+j0$. When $|K_a|=0$, there is no control, the system loop is closed: a self-oscillation pressure can occur. If $|K_a|=k$, perfect active cancellation of the acoustic feedback occurs; the loop is open because $\Sigma q_0(\omega)=0$. However, it is not necessary to perfectly compensate the acoustic feedback in order to reach the cancellation of the self-oscillation. According to the Nyquist criterion, the system is stable if the curve $G(\omega)$ does not encircle the critical point; in this case, there is not enough acoustic feedback to ensure self-

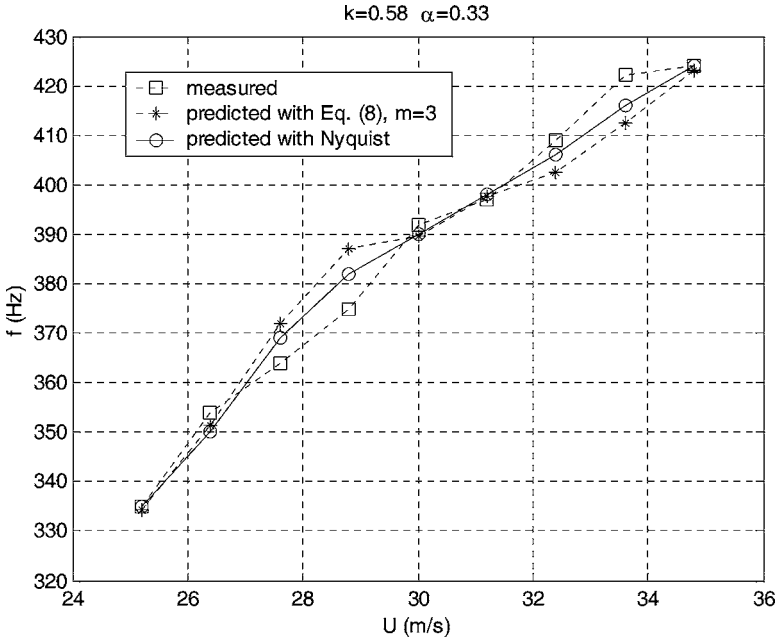


FIG. 4. Frequency of the self-sustained pressure fluctuation vs the free-stream velocity.

oscillation. The limit of stability is reached for the critical value:

$$\left| K_a^c \right| = k - \frac{1}{|G_b(\omega_m)|}, \tag{19}$$

which verifies that the curve $G(\omega)$ passes thought the critical point. Consequently, a small but sufficient feedback gain $|K_a| < |K_a^c|$ can cancel the existence of a self-oscillation.

IV. EXPERIMENTS

A. Experimental setup

Experiments were carried out in the 1.68 m diameter test section of an Eiffel-type wind tunnel. The free-stream velocity in the empty test section is up to 31 m/s, with a turbulence level of 1% in the 0–1000 Hz band. At this velocity, the integral length scale of the free-stream is around 17 mm for the streamwise velocity component and around 5.7 mm for the transverse velocity component. Because of the thick boundary layer developing on the wind tunnel walls, a wing-shaped model had been designed to hold the cavity. The model is 0.8 m wide and 1.74 m long, with a maximum thickness of 0.2 m, and it is based on the Joukowsky profile. A Pitot tube placed at the nose of the wing provides a velocity reference during pressure measurements. The presence of the model reduces the test section of the wind tunnel and thus increases maximum velocity above the cavity to 37 m/s (Mach 0.11). The width of the cavity is 0.8 m, and both the length and the depth are $L=150$ mm. The dimensions of the cavity were defined with reference to previous analytical and experimental results in order to obtain self-sustained oscillations of the mixing layer at low Strouhal numbers.[10]

The vibrating surface is an aluminum beam (3 cm $\times$ 30 cm) located along the downstream edge and attached to a shaker. To baffle the beam and to damp its natural vibration modes, a viscoelastic material is bonded between the beam circumference and the cavity. An accelerometer placed on the vibrating surface is used to measure its acceleration at different points. Six measurements along the vibrating surface showed that the acceleration level was roughly uniform in amplitude along the surface in the 200–450 Hz frequency band. Moreover, the vibration was roughly in phase along the beam. Hence, we assume that the beam vibrates in plane mode for frequencies below 450 Hz and that an accelerometer located at the center is sufficient to capture its dynamic.

Wall-mounted microphones placed at the bottom of the cavity were used to discern the self-sustained pressure fluctuations from the turbulence noise. Data were recorded during smooth acceleration ramps of the mean flow velocity (5–37 m/s in 6 mn) at a sampling frequency of 10 kHz. The spectrum evolution with the velocity of the flow was then calculated using Fourier transforms with 50% overlapping samples of 10 000 points, leading to a frequency resolution of 1 Hz. When the mean flow velocity increases from 25 to 36 m/s, a self-oscillation close to 400 Hz occurs. After 36 m/s, it disappears, and the pressure level of another tone close to 600 Hz dominates the response. Due to the validation of the actuator and its model for the 200–420 Hz frequency range, the study is focused on the self-oscillation nearer to 400 Hz.

The signal processing tools and the controller were developed with MATLAB/SIMULINK and implemented in a rapid prototyping dSpace system. The frequency sampling is 10 kHz. The low-pass filter f was realized with two second-order Butterworth ARMA filters characterized by a 200 Hz frequency band.

The frequency response functions (FRF) from the vibrating surface to three microphones located in the bottom of the cavity is measured on the 200–800 Hz frequency range. It appears that the FRF are roughly the same in amplitude and phase, below 425 Hz. This means that the cavity pressure response can be modelized by a one-dimensional system below 425 Hz. Hence, the FRF to the central microphone (the

one used for the active control) is the acoustic backward transfer function $G_b(\omega)$ used in this paper for the validated frequency domain: below 425 Hz.

B. Estimation of the parameters

An estimation of the parameters (α, κ, and k) can be obtained from the measurement of the frequency of the self-oscillation versus the free-stream velocity. For a 25 to 36 m/s free-stream velocity range, Fig. 6 presents the frequency of self-oscillation versus the free-stream velocity. The experimental frequency value increases linearly from 335 to 425 Hz. Consequently, and according to Eq. (8), the parameters $\alpha=0.33$ and $\kappa=0.58$ are identified. Figure 4 presents a good fit between the prediction of the frequency with Strouhal number relation, Eq. (1), by considering the stage $m=3$. Moreover, the value of $\kappa=0.58$ is consistent with the measure of the vorticity layer obtained with a PIV system.[13]

In order to estimate the value of the linear gain k, the describing function approach can be used. In this case, there is no assumption about the stage. Figure 5 presents the case of a 31.2 m/s velocity to illustrate the approach. The Nyquist plot shows that the self-oscillation can occur at 398 Hz, which compares favorably with the frequency measured at 397 Hz (Fig. 4). To predict the self-oscillation of the pressure the point $-1/k$ must be encircled by the Nyquist curve $G(\omega)$. Moreover, this encirclement must be valid from 27.6 to 32.4 ms^{-1}. To respect this constraint, the point must be roughly located at $-1/k \approx -1.5\times10^{4}$ Pa s m^{-3}; hence, the gain of the acoustic feedback must be $k=6.7\times10^{-5}$ $\mathrm{m}^3\,\mathrm{s}^{-1}\,\mathrm{Pa}^{-1}$. The predicted frequencies obtained with this approach are presented in Fig. 4; they compare favorably to the experimental ones.

C. Estimation of the volume velocity in open loop control

For open loop control, forced frequency (the frequency of the actuator) must be very close to natural self-oscillation

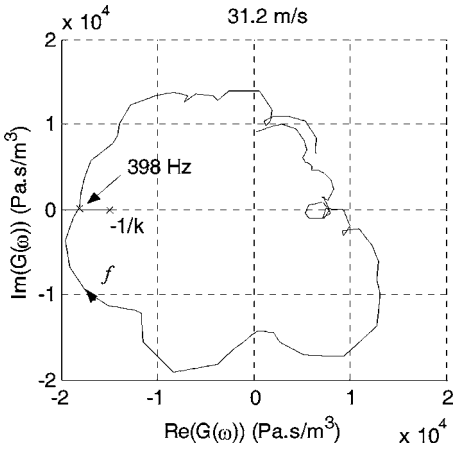


FIG. 5. Nyquist plot of $G(\omega)$ for a free-stream velocity $U_\infty=31.2$ m s^{-1}: the describing function analysis leads to predict a self-oscillation at $f_0=398$ Hz.

frequency (±1 Hz) for synchronization. Figure 6 shows the pressure amplitude versus the volume velocity of the actuator, $|q_1|$. Without control ($|q_1|=0$), the pressure level is 18 Pa; the natural self-oscillation is not synchronized with the forced frequency at 398 Hz. When the volume velocity amplitude $|q_1|$ is increased, the pressure level increases. When $|q_1|>0.5\times10^{-4}$ $\mathrm{m}^3\,\mathrm{s}^{-1}$, the pressure oscillation is perfectly synchronized with the actuator: there is a −1.8 rad phase shift between them (with a standard deviation of 0.3 rad). This shows that an increase of the volume velocity amplitude leads to an evolution from a natural self-oscillating mode to a synchronized oscillation mode at the forced frequency. The estimated value of the saturated volume velocity is $Q_{\max}=10^{-3}$ $\mathrm{m}^3\,\mathrm{s}^{-1}$. The predicted amplitude calculated with Eq. (12) compare favorably to the experimental ones presented in Fig. 6.

D. Optimal phase lag in closed loop control

The closed loop is considered with a saturated command such that volume velocity amplitude is a constant value. The

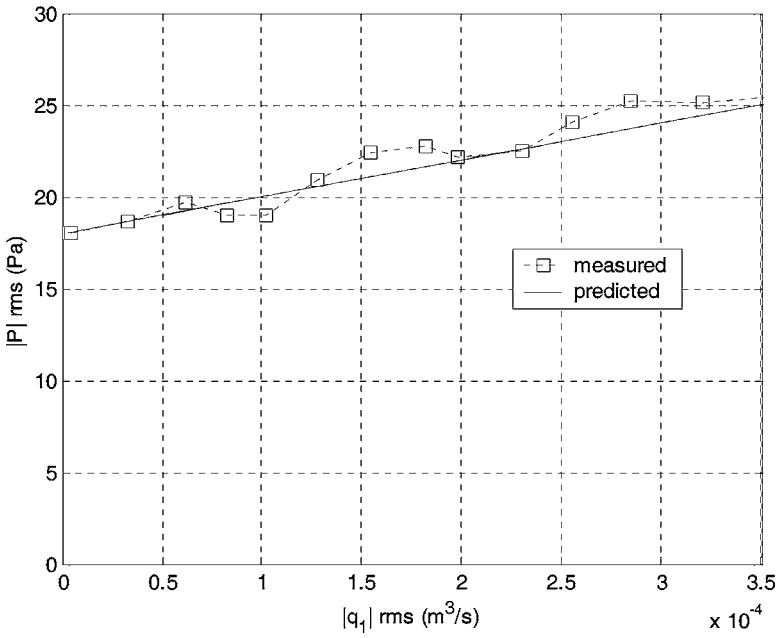


FIG. 6. Pressure levels versus the volume velocity for forced oscillation at 398 Hz.

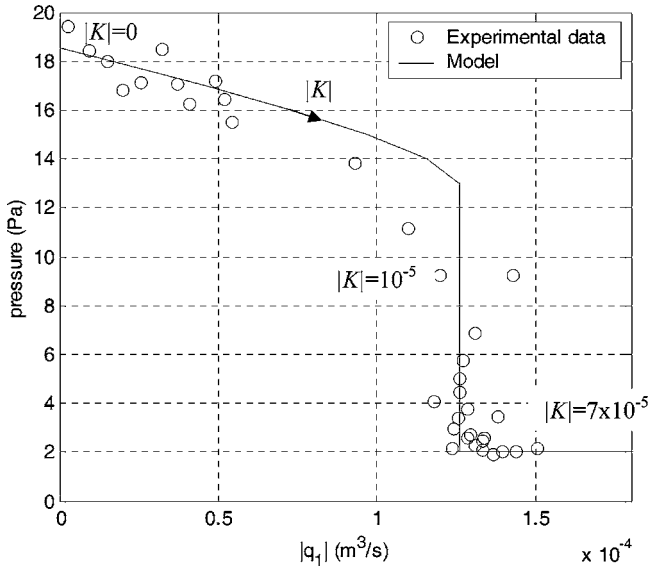


FIG. 7. Pressure amplitude versus the volume velocity of the actuator when the modulus gain controller, $|K|$, is tuned from 0 to 7×10^{-5} m^3 s^{-1} Pa^{-1}.

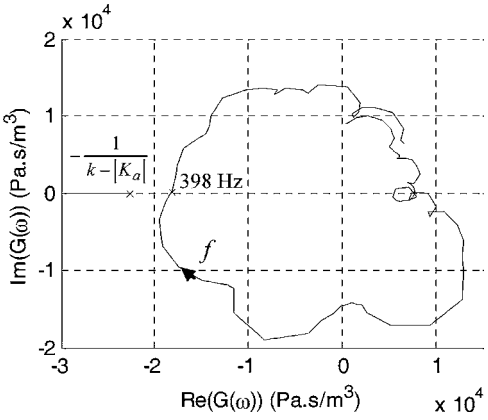


FIG. 8. Nyquist plot of $G(j\omega)$ (31 m/s) when the describing function analysis leads to predict the system stable with the gain $K_a=2K_a^c$.

tuning parameter is the phase lag of the feedback gain K. It is tuned from 0 to 6 rad, in order to measure its impact on the closed loop system. For a value close to 5 rad, the amplification is maximal. In this case, the active control reinforces the instability because it acts as a positive feedback. For a value close to 1.7 rad, the maximal attenuation is reached. It is interesting to remark that this optimal phase lag value of K is roughly the value of the phase lag of the system in open loop control minus π (1.7 rad-π=−1.8 rad); hence, the maximal attenuation is reached when the phase lag of the controller is tuned to obtain the perfect negative feedback. It is important to remark that the phase-lag value of the open loop control must be considered and not the phase lag of the acoustic backward transfer function. This result is typical, and is obtained with different configurations.

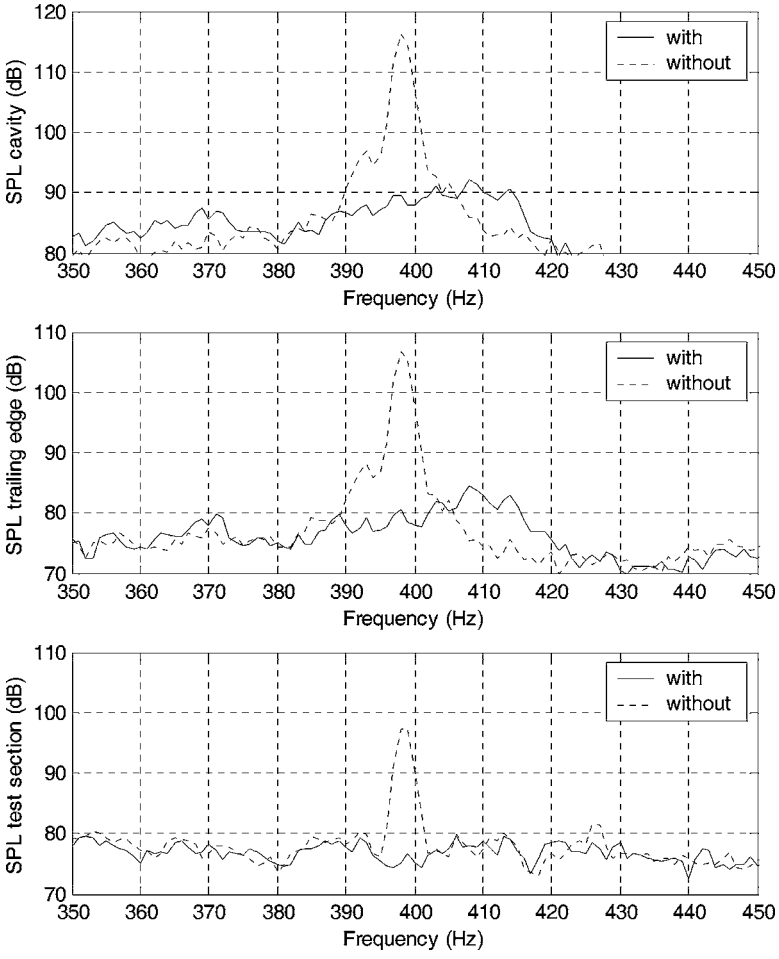


FIG. 9. Sound pressure level with and without active control measured in the cavity (top), at the trailing edge (middle), and at 1 m from the cavity in the test section (bottom).

E. Critical feedback gain

Because the measured transfer function at 398 Hz is $|G_b(\omega_m)| = 1.8 \times 10^4$ s Pa m^{-3} and the estimated value of the acoustic feedback is $k = 6.7 \times 10^{-5}$ m^3 s^{-1} Pa^{-1}, the predicted critical gain with Eq. (19) is $|K_a^c| = 1.14 \times 10^{-5}$ m^3 s^{-1} Pa^{-1}. In order to validate this prediction, the linear gain feedback is tuned from 0 m^3 s^{-1} Pa^{-1} to 7×10^{-5} m^3 s^{-1} Pa^{-1} (with the optimal phase lag). Figure 7 shows the pressure amplitude, $|p|$, versus the volume velocity of the actuator, $|q_1|$, for this experiment.

For low values of the feedback gain, below 1×10^{-5} m^3 s^{-1} Pa^{-1}, the volume velocity amplitude increases linearly and the pressure fluctuation is attenuated by classical linear interference. When critical gain is reached, close to 1×10^{-5} m^3 s^{-1} Pa^{-1}, with a volume velocity close 1×10^{-4} m^3 s^{-1}, stability limit is reached. Over this value, the pressure is decreased to the ground level in the subband (2 Pa). Self-oscillation is attenuated and, instead, the feedback gain is increased. For a large gain (superior than 5×10^{-5} m^3 s^{-1}), the command amplitude is always saturated and the active control is still efficient. Figure 8 presents the interpretation of this result: when the feedback gain increases, the location of the critical point moves to the left of the curve $G(\omega)$; hence, the unstable physical system becomes a stable one.

Figure 9 presents the spectrum with and without active control when $|K| = 7 \times 10^{-5}$ ($|q_1| = 1.5 \times 10^{-4}$ m^3/s). The tone is completely attenuated to the ground noise (25 dB of attenuation). A microphone located at the trailing edge also shows a 25 dB attenuation to the ground level. A microphone located in the test section (1 m from the cavity) also shows a ground noise reduction (17 dB attenuation due to ground noise). We can conclude that the aeroacoustic source is controlled and, consequently, the unwanted radiated noise.

To validate that this active control is efficient for different frequencies, active control was performed at different speeds from 25 to 35 m/s. Without active control, the power spectrum (in the frequency band $[f_0 - \Delta f; f_0 + \Delta f]$ with $\Delta f = 25$ Hz) progressively increases from 96 dB at 25 m/s (335 Hz) to 118 dB at 31.2 m/s (397 Hz) and, after, it decreases to 105 dB at 35 m/s (424 Hz). The maximum level corresponds to the acoustic reinforcement close to 396 Hz. With active control, the tone is roughly attenuated to the ground noise; the power spectrum in the frequency band is 94 dB at 25 m/s, 105 dB at 31.2 m/s and 102 dB at 35 m/s.

V. CONCLUSIONS

The describing function analysis was presented to analyze the active control of self-sustained pressure fluctuations. This approach was applied to a large shallow cavity in a wing-shaped model where sound pressure was generated by the mixing layer instability. It was shown that a vibrating surface located at the trailing edge can be used to control the self-oscillation: to amplify the pressure in an open loop, to attenuate it in a closed loop. A validation of the describing function analysis was obtained through a comparison between the theoretical prediction and experimental data. The main point of this analysis is to consider the acoustic backward transfer function, without simplification: the whole acoustical environment of the cavity was considered. By doing so, the proposed method is therefore expected to be efficient in practical implementations.

With the proposed model and the original device, two main parameters were estimated. In an open loop, adjusting the volume velocity generated by the vibrating surface allows us to observe the amplitude saturation of the instability and, consequently, estimate its equivalent volume velocity amplitude. In a closed loop with the perfect phase-lag compensation, adjusting the feedback gain of the narrow-band controller makes it possible to estimate the acoustic feedback gain between the pressure fluctuation in the cavity and the volume velocity of the aeroacoustic source. Works are currently in progress to insert the retroaction model approach exposed here into an analytical model of the mixing layer.

ACKNOWLEDGMENTS

This work was financially supported by the "Conseil Régional du Poitou-Charentes" (France) and partially supported by the "Groupe De Recherche Mécanique des Fluides Actives" (CNRS, France). P. Micheau would like to gratefully thank the Laboratoire d'Études Aérodynamiques, Poitiers, France, where all this research was developed.

[1] V. Sarohia and P. F. Massier, "Control of cavity noise," AIAA J. **14**, 833–837 (1977).
[2] L. N. Catafestta, D. R. Williams, C. W. Rowley, and F. Alvi, "Review of active control of flow-induced cavity resonance," *AIAA Fluid Dynamics Conference*, 23–26 June, Orlando, 2003, Vol. **33**, p. 3567.
[3] P. A. Nelson and S. J. Elliot, *Active Control of Sound* (Academic, New York, 1992).
[4] F. T. Calkins and D. J. Clingman, "Vibrating surface actuators for active flow control," Proc. SPIE **4698**, 85–96 (2002).
[5] S. Ziada, H. Ng, and C. E. Blake, "Flow excited resonance of a confined shallow cavity in low Mach number flow and its control," J. Fluids Struct. **18**, 79–92 (2003).
[6] H. Kook, L. Mongeau, and M. A. Franchek, "Active control of pressure fluctuations due to flow over Helmholtz resonators," J. Sound Vib. **255**, 61–76 (2002).
[7] L. Chatellier, "Modélisation et contrôle actif des instabilités aéroacoustiques en cavité sous écoulement affleurant, Ph.D. thesis, Université de Poitiers, September 2002.
[8] J. E. Rossiter, "The effect of cavities on the buffeting of aircraft," *Royal Aircraft Establishment Technical Memorandum*, 1962, p. 754.
[9] M. S. Howe, "Low Strouhal number instabilities of flow over apertures and wall cavities," J. Acoust. Soc. Am. **102**, 772–780 (1997).
[10] L. Chatellier, J. Laumonier, and Y. Gervais, "Theoretical and experimental investigations of low Mach number turbulent cavity flows," Exp. Fluids **36**, 728–740 (2004).
[11] X. Slotine and X. Li, *Applied Nonlinear Control* (Prentice–Hall, Englewood Cliffs, NJ, 1991).
[12] T. D. Mast and A. D. Pierce, "Describing-function theory for flow excitation of resonators," J. Acoust. Soc. Am. **97**, 163–172 (1995).
[13] H. Kook and L. Mongeau, "Analysis of the periodic pressure fluctuations induced by flow over a cavity," J. Sound Vib. **251**, 823–846 (2002).
[14] P. Micheau and P. Coirault, "Adaptive controller using filter banks to reject multi-sinusoidal disturbance," Automatica **36**, 1659–1664 (2000).
[15] P. Micheau, P. Coirault, L. Hardouin, and J. Tartarin, "Adaptive rejection of the preponderant harmonic of a pulsed flow," IEEE Trans. Control Syst. Technol. **4**, 452–458 (1996).
[16] P. Micheau and S. Renault, "Active control of the complex envelope associated with a low damped mode," Mech. Syst. Signal Process. **20**(3), 646–661 (2006).

On the use of a diffusion equation for room-acoustic prediction

V. Valeau[a)]
LEPTAB, Université de La Rochelle, Pôle Sciences et Technologies, Avenue M. Crépeau, 17042, La Rochelle Cedex 01, France

J. Picaut
Laboratoire Central des Ponts et Chaussées, Section Acoustique Routière et Urbaine, Route de Bouaye, Boîte Postale 4129, 44341 Bouguenais Cedex, France

M. Hodgson
University of British Columbia, Acoustics & Noise Research Group, SOEH and MECH, 3rd Floor, 2206 East Mall, Vancouver, BC, Canada V6T 1Z3

(Received 10 August 2005; revised 29 November 2005; accepted 2 December 2005)

This paper presents an alternative model for predicting the reverberant sound field in empty rooms with diffusely reflecting boundaries, based on the generalization and the numerical implementation of a diffusion equation for the energy density. The paper focuses on the source term and the boundary conditions of the diffusion equation, both for the steady state and the time-varying state, in order to make computational use of the model. In addition, theoretical analysis of the diffusion equation shows that the diffusion model may be considered as an extension of the classical theory of reverberation to nondiffuse sound fields. The numerical model is first applied to a cubic room and shows a very good agreement with statistical theory. Two numerical applications are also given for a long room and a flat room; results are in good agreement with numerical results from a ray-tracing software. The main advantage of the present model is its capability to be applied regardless of the complexity of the room shape, and that it gives results at any receiver location, with a low calculation time. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2161433]

PACS number(s): 43.55.Br, 43.55.Ka, 43.55.Cs [NX] Pages: 1504–1513

I. INTRODUCTION

Sound-field modeling in room acoustics has attracted considerable attention in recent decades. Many developments have been proposed to predict the sound-field distribution and the sound decay in rooms. In particular, Sabine's statistical theory for diffuse sound fields has been widely used to estimate the sound decay and the steady-state reverberant field in quasicubic rooms.[1,2] For corridors and flat rooms—i.e., rooms with small height compared to their width and length—for which the diffuse-sound-field theory cannot be applied, many mathematical models have been derived to predict the sound-level distribution and the spatial variations of reverberation time.[3] For real rooms—i.e., with complex shapes and with partially diffusely reflecting surfaces—computer methods such as the ray-tracing model have been developed and are now very popular. However, by increasing the complexity of the room shape, the computation time increases dramatically and can be several hours for engineering accuracy.

In this paper, we propose an alternative approach to predict the sound field in arbitrary rooms with a satisfactory accuracy and with low computation time. This model, based on a physical analogy with the diffusion of particles in a scattering medium, was recently derived to predict the spatial variations of the reverberant sound field in long rooms[4,5] and in urban areas[6] with diffusely reflecting boundaries, for both steady-state and time-varying sources. However, no absolute sound levels were obtained and the terms accounting for energy losses at boundaries were only preliminarily analyzed. The applications to long rooms involved purely analytical, one-dimensional analyses. The present study proposes a generalization of this theory in the case of arbitrary, three-dimensional enclosures, in order to achieve a computational use of the diffusion equation in the case of room-acoustic applications.

First, the acoustic diffusion model is presented in Sec. II. To extend the use of this theory in the case of arbitrary rooms, some major points are then discussed in Sec. III, as follows: (i) the definition of *ad hoc* source terms; (ii) the development of two models with different treatments of the energy absorption at wall surfaces, and the discussion of their equivalence; and (iii) the link of the present acoustic diffusion theory with the classical diffuse sound field theory. Finally, the numerical implementation of the model, based on the finite-element method, is presented in Sec. IV with three examples: a room with "proportionate dimensions," a long room, and a flat room. Sound predictions are then compared with results obtained using a ray-tracing model.

II. ANALOGY WITH THE DIFFUSION OF PARTICLES BY SCATTERERS

The model presented in this paper is based on the analogy of the sound energy density with a density of "sound

[a)]Present address: LEA, UMR CNRS 6609, Bâtiment K, 40 Avenue Recteur du Pineau, 86022 Poitiers Cedex, France. Electronic mail: vvaleau@univ-poitiers.fr

particles"[7] traveling at velocity c along straight lines. For further use, the case of a source in an infinite medium is first briefly considered. Let us consider a source of infinitely small dimension emitting q particles per second omnidirectionally. In a time interval dt, this source emits qdt particles. After a time t, all particles are located within a spherical shell of thickness cdt and radius ct, with enclosed volume $4\pi(ct)^2cdt$. The concentration of particles $w(r)$ at a distance r from the source is then

$$w(r=ct)=\frac{qdt}{4\pi(ct)^2cdt}=\frac{q}{c4\pi r^2}. \tag{1}$$

The acoustic energy density at a distance r from a source of power P emitting omnidirectionally is $P/(c4\pi r^2)$, c being the sound speed. The analogy with the particle density given by Eq. (1) is then straightforward for a source in an infinite medium, the quantity q being analogous to the acoustical power P.

The present study uses a model initially derived by Ollendorff[8] and further developed by Picaut *et al.*,[4] for modeling the acoustics of enclosed spaces with diffusely reflecting surfaces, based on an analogy with the model for the diffusion of particles in a medium containing spherical scattering obstacles, as presented by Morse and Feshbach.[9] Following their work, let us consider a room defined by its volume V, its surface area S, and an average surface absorption coefficient $\bar{\alpha}$. It is possible to consider the scattering of sound particles by the room surfaces as analogous to the scattering of the same particles by a uniform distribution of scatterers throughout the volume. The probability $\mathcal{P}(x)$ that a particle has covered a distance x without any collision with a scatterer is assumed to be described by an exponential distribution,[9]

$$\mathcal{P}(x)=Qne^{-Qnx}, \tag{2}$$

where n is the density of scatterers—i.e., the number of scattering particles per volume unit—and Q the particle scattering cross-section. $\bar{\alpha}$ (between 0 and 1) is the probability for a scatterer to absorb a particle after a collision. The mean free path λ between two collisions is

$$\lambda=\int_0^\infty x\mathcal{P}(x)dx=\frac{1}{Qn}. \tag{3}$$

In room acoustics, a common analytical expression for the mean free path in a room is

$$\lambda=4V/S. \tag{4}$$

Many authors[2,7,10] have discussed extensively the validity of this relation. We shall keep in mind here that, in particular, this expression is valid in the case of a room with diffusively reflecting boundaries—i.e., with directional reflection properties that are described by Lambert's law.[2]

In the present analogy, Picaut *et al.*[4] set the value for the mean free path between two collisions for a particle to $\lambda=4V/S$. Morse and Feshbach[9] show that, given the slow rates of change with time involved in diffusion phenomena, the local particle-density flux $\mathbf{J}(\mathbf{r},t)$ can be approximated as the gradient of the particle density,

$$\mathbf{J}(\mathbf{r},t)=-D\nabla w(\mathbf{r},t). \tag{5}$$

Under this assumption, the particle density can be shown to be the solution of a diffusion equation. The analogy with room acoustics allows the assumption that the acoustic energy density in the room at a position $\mathbf{r}$ and time t is governed by the same diffusion equation,

$$\frac{\partial w(\mathbf{r},t)}{\partial t}-D\nabla^2 w(\mathbf{r},t)+\sigma w(\mathbf{r},t)=0, \tag{6}$$

where ∇^2 is the Laplace operator. D is the so-called diffusion coefficient, and the term σ accounts for absorption at the room boundaries, defined by a mean room-surface absorption coefficient $\bar{\alpha}$. Their analytical expressions are taken directly from the theory of diffusion for particles in a scattering medium,[9]

$$D=\frac{\lambda c}{3}, \tag{7}$$

$$\sigma=\frac{\bar{\alpha}c}{\lambda}. \tag{8}$$

The diffusion coefficient D used here is different from a usual definition in room acoustics, which is the proportion of diffuse reflections at the room surfaces.[11] The diffusion coefficient D as described in Eq. (7) is a term that takes into account the room morphology through its mean free path.

III. MODEL FOR ROOM-ACOUSTIC PREDICTIONS

With the objective of making computational use of the diffusion equation for three-dimensional room-acoustic predictions, this section focuses on the source term in the diffusion equation and on the definition of the boundary conditions. Finally, it clarifies the link between the present theory and the classical theory of diffuse sound fields.

A. Source models

1. Volume source

Let us consider an omnidirectional sound source radiating a sound power P. The analogy for the direct field presented at the beginning of Sec. II shows that a source of constant power P is analogous to a point source emitting q particles per second. In the theory of the diffusion of particles by scatterers, a point source emitting q particles per second is properly modeled by a source term equal to $q\delta(\mathbf{r}-\mathbf{r}_s)$,[9] where $\mathbf{r}_s$ denotes the position of the source. In an analogous way, for a room-acoustic problem, the right-hand source term in the diffusion equation, which models adequately a point source with arbitrary acoustic power $P(t)$ located at $\mathbf{r}_s$, is

$$\frac{\partial w(\mathbf{r},t)}{\partial t}-D\nabla^2 w(\mathbf{r},t)+\sigma w(\mathbf{r},t)=P(t)\delta(\mathbf{r}-\mathbf{r}_s). \tag{9}$$

As for all other terms in this differential equation, this source term has the dimension of a power density. The case of a source occupying a subdomain V_s of volume v and radiating

the power $P(t)$ can be modeled in a similar way by taking the following source term $G(\mathbf{r},t)$:

$$G(\mathbf{r},t) = \frac{P(t)}{v} f_s(\mathbf{r}), \tag{10}$$

the function $f_s(\mathbf{r})$ being equal to 1 inside V_s and to 0 otherwise. The integration of $G(\mathbf{r},t)$ over the subdomain V_s must be equal to the total power output of the source $P(t)$, and the above expression meets this condition. This model will be preferred for performing numerical simulations (Sec. IV) because it diminishes the singularity of the Green's function of the diffusion equation.

If the point source is impulsive, the diffusion equation becomes

$$\frac{\partial w(\mathbf{r},t)}{\partial t} - D\nabla^2 w(\mathbf{r},t) + \sigma w(\mathbf{r},t) = E_0 \delta(\mathbf{r}-\mathbf{r}_s)\delta(t-t_0), \tag{11}$$

where E_0 is the energy produced by the source at time t_0. For practical purposes, a source emitting a constant power P in a short time interval Δt can be considered. In this case, E_0 can be approximated by $E_0 \simeq P\Delta t$.

2. Surface source

The transmission of sound through a room boundary can also be modeled by setting an appropriate boundary condition. Suppose that some sound-energy transmission is possible through a portion S of the room surfaces. If $\mathbf{n}$ is the local outgoing vector normal to the wall, and J_{in} is the ingoing local energy flux due to the transmission, then the boundary condition that takes the transmission into account will be

$$\mathbf{J}\cdot\mathbf{n} = -D\frac{\partial w}{\partial n} = -J_{\text{in}} \quad \text{on } S, \tag{12}$$

the negative sign accounting for the ingoing nature of the transmitted energy. A sample application will be presented in Sec. IV.

B. Boundary conditions

1. Homogeneous Neumann boundary conditions

The model described in Sec. II is derived for an infinite scattering medium. The analytical solution of the one-dimensional form of the diffusion equation of Eq. (6) has been used successfully for calculating the sound attenuation and the reverberation time in infinitely long rooms.[4,5] Indeed, under the reasonable assumption that the energy density is constant over a cross section of a long room, an infinitely long room can be considered as analogous to a one-dimensional scattering medium.

However, to make possible the use of a three-dimensional diffusion process for an arbitrary room shape, the equivalent scattering medium has to be bounded in some way by setting up *ad hoc* boundary conditions. The absorption of the room surfaces are taken into account in the diffusion equation of Eq. (6) by the term $\sigma w(\mathbf{r},t)$, with $\sigma = \bar{\alpha}c/\lambda$. This term expresses the loss of energy in the room per unit volume and per unit time due to the mean room-absorption coefficient $\bar{\alpha}$. Hence the boundary conditions need not involve any extra energy absorption. This can be modeled appropriately by homogeneous Neumann boundary conditions,[9] which impose no energy flux through the room surfaces. The reflection of sound particles by scatters (or, equivalently, by surfaces) is still entirely described by the local diffusion equation (6) and the mean free path [Eq. (4)]. The homogeneous boundary conditions ensure that the sound particles cannot "escape" from the room volume. The diffusion model for both volume and boundaries should then be

$$\frac{\partial w(\mathbf{r},t)}{\partial t} - D\nabla^2 w(\mathbf{r},t) + \sigma w(\mathbf{r},t) = F(\mathbf{r},t) \quad \text{in } V, \tag{13}$$

with

$$\mathbf{J}(\mathbf{r},t)\cdot\mathbf{n} = -D\,\nabla w(\mathbf{r},t)\cdot\mathbf{n} = 0, \quad \text{i.e.,}$$
$$\frac{\partial w}{\partial n} = 0 \quad \text{on } \partial V, \tag{14}$$

where ∂V denotes the room boundaries and $F(\mathbf{r},t)$ an arbitrary acoustic source term. The solutions for the set of Eqs. (13) and (14), referred to as Model 1, will be further investigated analytically in Sec. III C, and numerically in Sec. IV.

2. Mixed boundary conditions

The model described in the last section is limited in the sense that the room absorption is taken into account through its mean value $\bar{\alpha}$. Some difficulties can arise to model the local effects on the sound field induced by absorbing materials, for instance.[12] To overcome this problem, different boundary conditions could be set up that take into account the distribution of the room absorption. A common way of describing energy exchanges at a boundary is by the use of a mixed boundary condition, in the following form:

$$\mathbf{J}(\mathbf{r},t)\cdot\mathbf{n} = -D\frac{\partial w}{\partial n} = hw(\mathbf{r},t) \quad \text{on } \partial V, \tag{15}$$

where h is the so-called exchange coefficient. Under this condition (for instance, as widely used in heat transfer),[13] the local energy flux outside the room is assumed to be proportional to the local energy density. At the same time, this boundary condition involves some energy absorption within the room volume, as it models an energy loss per unit area and per unit time. Hence, the local differential equation as written in Eq. (13) should be modified to remove the absorption term $\sigma w(\mathbf{r},t)$. As a consequence, the boundary condition of Eq. (15) has to be equivalent to this former term in terms of energy loss for the room. The term h should be chosen so that the energy flux through the room boundaries equals the absorption over the whole room due to the term $\sigma w(\mathbf{r},t)$, i.e.,

$$\int_V \sigma w(\mathbf{r},t)dV = \int_{\partial V} h(S)w(\mathbf{r},t)dS. \tag{16}$$

If the sound field is assumed to be totally diffuse, i.e., the energy density $w(\mathbf{r},t)$ is uniform $\forall\mathbf{r}$, the energy density can be taken out of the integrals. Then, by using Eqs. (8) and (4), Eq. (16) leads simply to

$$\frac{c\bar{\alpha}S}{4} = \int_{\partial V} h(S)dS. \tag{17}$$

If the room walls are made of N surfaces S_i with constant absorption α_i (i.e., $\bar{\alpha}=\sum_{i=1}^{N}\alpha_i S_i/S$ for $i=1,\ldots,N$), then one can attribute an exchange coefficient h_i to each surface portion so that Eq. (17) can be written

$$h_i = \frac{c\alpha_i}{4}, \quad i = 1, \ldots, N. \tag{18}$$

The energy flux through an infinitely small surface portion dS of absorption coefficient α is then

$$\mathbf{J}\cdot\mathbf{n}\, dS = \frac{cw(\mathbf{r},t)}{4}\alpha\, dS, \tag{19}$$

which is in accordance with diffuse sound-field theory.[2]

Relation (18) has been derived for the case of an uniform sound field but, for rooms with special shapes, such as long rooms, for which the acoustic energy density is known not to be uniform in the long dimension,[3] the same result holds (as shown in Ref. 5 in a somewhat different way than below). For instance, let us consider an infinitely small portion of a long room of length dx and volume sdx. x is the axis of the longer dimension of the room, s its cross-sectional area in m^2, and L its length. It is reasonable to suppose that the energy density w varies only along x, i.e., it is constant over cross-section s. Let us denote p the perimeter of the cross-section, and α the absorption coefficient of the side walls. Absorption could be different for each side wall without difficulty—however, the differences in the absorption coefficients of the side walls should not be too large, so that the one-dimensional assumption remains acceptable. For such a long room, if the area of the hall extremities is negligible compared to the side-wall area, the mean-free-path expression $\lambda=4V/S$ can be approximated by $4s/p$ (as $V\simeq sL$ and $S\simeq pL$). Equation (16) can then be written

$$\frac{c\alpha p}{4s}w(x,t)sdx = hw(x,t)pdx, \tag{20}$$

which immediately leads to $h=c\alpha/4$, in accordance with Eq. (18). Hence, in the case of a long room with sound field only varying along the longer dimension of the room, the boundary condition (15) with $h=c\alpha/4$ is close to being equivalent (given an approximation on the mean free path λ) to the local term $\sigma w(\mathbf{r},t)$ in terms of energy absorption. Similar deductions can also be extended to the case of a flat room. Hence, for room shapes such as in long or flat rooms, leading to substantially non-uniform sound fields, the boundary condition Eq. (19) which is, strictly speaking, valid for diffuse sound fields, is a good approximation for modeling energy losses at the room surfaces.

In conclusion, to use a diffusion process in a room, the following system of equations is solved:

$$\frac{\partial w(\mathbf{r},t)}{\partial t} - D\nabla^2 w(\mathbf{r},t) = F(\mathbf{r},t) \quad \text{in } V, \tag{21}$$

$$D\frac{\partial w(\mathbf{r},t)}{\partial n} + \frac{c\alpha}{4}w(\mathbf{r},t) = 0 \quad \text{on } \partial V. \tag{22}$$

This set of equations will be referred to as Model 2. This model is analogous to the model for heat transfer within a heat-conductive medium (i.e., the room) with heat exchanges at the medium boundary (i.e., the room surfaces). For a room with a sound field close to being uniform, both Models 1 and 2 are equivalent. For rooms with shapes usually leading to significant spatial variations in the sound field (such as rooms with long or flat shapes), this is also valid. Section IV shows that the solutions of both systems are indeed very similar. The advantage of the second formulation is that the boundary conditions are directly dependent on the local boundary absorption coefficient, and so it can handle arbitrary variations of absorption. Conversely, in the first formulation, only a mean absorption coefficient $\bar{\alpha}$ is introduced in the model.

The frequency dependence of the phenomena are taken into account through the frequency dependence of the absorption coefficients of the room boundaries. Hence, for a calculation of the sound field in a given frequency band, the sets of equations for Models 1 or 2 must be solved with the appropriate values of α for this band. The frequency-dependent terms for each model are $\sigma w(\mathbf{r},t)$ for Model 1, and the boundary condition [Eq. (22)] for Model 2.

C. Relation with diffuse sound-field theory

1. Room energy balance

The equations describing the acoustic energy density w for both the classical diffuse sound-field theory (called *statistical theory* in the following) and the present diffusion analogy are now briefly compared. One main assumption for a diffuse sound field being that the quantity w is uniform throughout the room, no local differential equation for w can be derived. However the energy balance for a room of volume V containing a sound source of power output $P(t)$ can be written[2]

$$V\frac{dw(t)}{dt} + V\frac{c\bar{\alpha}S}{4V}w(t) = P(t). \tag{23}$$

This can be compared to the energy balance for the same room given by the diffusion theory. For the first diffusion model (Model 1) of Sec. III B 1, this balance is achieved by integrating Eq. (13) over the room volume V and replacing σ with $c\bar{\alpha}/\lambda$ (with $\lambda=4V/S$),

$$\int_V \frac{\partial w(\mathbf{r},t)}{\partial t}dV + \int_V \frac{c\bar{\alpha}S}{4V}w(\mathbf{r},t)dV - D\int_V \nabla^2 w(\mathbf{r},t)dV = \int_V F(\mathbf{r},t)dV. \tag{24}$$

By using Gauss' theorem[9] and the boundary condition of Eq. (14), the integral term containing the Laplacian operator ∇^2 turns out to be 0

$$D\int_V \nabla^2 w(\mathbf{r},t)dV = D\int_{\partial V} \frac{\partial w(\mathbf{r},t)}{\partial n} dS = 0, \tag{25}$$

so that the room energy balance for the first diffusion model becomes

$$\int_V \frac{\partial w(\mathbf{r},t)}{\partial t} dV + \int_V \frac{c\bar{\alpha}S}{4V} w(\mathbf{r},t)dV = P(t) \tag{26}$$

if the source term of Eq. (24) is properly set, as described in Sec. III A 1. Equations (23) and (26) are obviously similar, with the difference that the second equation is just an extension of the first (i.e., statistical theory) for a spatially varying sound field. For $w(\mathbf{r},t)$ constant in the room, Eq. (26) reduces to Eq. (23).

With the second model [Eqs. (21) and (22)], one can apply the same calculation steps: volume integration of the diffusion equation, and application of Green's theorem to the ∇^2 integral term together with boundary condition as stated in Eq. (22). The following room energy balance follows:

$$\int_V \frac{\partial w(\mathbf{r},t)}{\partial t} dV + \int_{\partial V} \frac{c\alpha}{4} w(\mathbf{r},t)dS = P(t), \tag{27}$$

which is equivalent to Eq. (26) on the conditions discussed in the previous section. Model 2 can then also be seen to be an extension of the statistical theory to non-uniform sound fields.

2. Steady-state sound field

For a constant source power P, the energy balance (23) for diffuse-field theory gives the well-known energy density

$$w_s = \frac{4P}{c\bar{\alpha}S}. \tag{28}$$

Equation (13) for a constant-power point source is no longer time-dependent if the steady state is reached,

$$D\nabla^2 w(\mathbf{r},t) - \sigma w(\mathbf{r},t) = -P\delta(\mathbf{r}-\mathbf{r}_s). \tag{29}$$

This is a Helmholtz equation for which the Green's function for an unbounded medium is[9]

$$G_u(\mathbf{r},\mathbf{r}_s) = \frac{Pe^{-\sqrt{(\sigma/D)}r}}{4\pi Dr}, \tag{30}$$

with $r=\|\mathbf{r}-\mathbf{r}_s\|$. In order to investigate the ability of the diffusion model to describe the steady state of a room, one can compute the analytical solution for Eq. (29) [i.e., a Green's function $G(\mathbf{r},\mathbf{r}_s)$ which satisfies the homogeneous Neumann boundary conditions (14)] by using the image-source method presented, for example, in Ref. 9, p. 812. For instance, let us consider a room of cubic shape $10\times10\times10$ m^3 with surfaces with absorption coefficient $\alpha=0.05$, containing a source of constant power P located at its geometrical center. A three-dimensional network of image sources is constructed, up to images of order 20. Each image is weighted by no absorption factor, as the boundary conditions are Neumann-homogeneous. The Green's function $G_u(\mathbf{r},\mathbf{r}_s)$ of the unbounded medium is computed at each image location. The relative deviation from statistical theory [Eq. (28)] for the obtained energy density, resulting from the combination of the actual source and the image sources, along a straight line passing through the source and parallel to one of the sides of the cube, is plotted on Fig. 1.

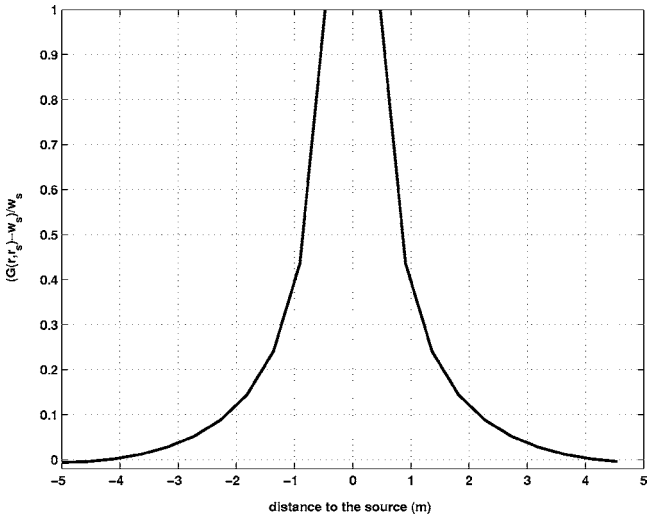


FIG. 1. Relative deviation from the statistical theory of the reverberant energy density calculated with the Green's function [Eq. (30)] and the image-source method, for a $10\times10\times10$ m^3 cubic room of uniform absorption coefficient 0.05. Results are plotted as a function of the distance r to the source, along a line parallel to one of the sides of the cube.

The singularity of function $G_u(\mathbf{r},\mathbf{r}_s)$ at the source location ($r=0$) is responsible for an infinite deviation of the modeled reverberant field from the statistical theory. Agreement becomes satisfactory (deviation of less than 10%) for distances to the source greater than about 2 m. The deviation is negligible near the surfaces. As pointed out by Morse and Feshbach (Ref. 9, p. 191), due to the approximation involved for obtaining the diffusion equation, the Green's function is invalid for distances to the source somewhat smaller than the room mean free path, where it has very large values due to the $1/r$ decrease for $r\to0$. In fact, this solution models an erroneous direct sound field, i.e., particles that, for small values of r, have a high probability [Eq. (2)] for not having hit a scatterer yet. For this reason, the singularity at the source location, predominant close to the source, should be limited in some way and replaced by the correct decrease in $1/r^2$. This will be discussed in the numerical results presented in Sec. IV. In addition, the analogy with a medium of scattering particles is questionable for distances smaller than the mean free path, as relation (2) means that there is some chance for a sound particle to hit a surface even at infinitely small distances from the source, which is obviously non-physical in the present case. However this apparent discrepancy does not prevent the diffusion model from giving satisfactory results if the problem is properly defined numerically.

3. Sound-energy decay

For a sound source which ceases to radiate at $t=t_0$, the solution of the energy balance in Eq. (23) is

$$w_s(t) = w_0 e^{-(c\bar{a}S/4V)(t-t_0)} = w_0 e^{-\sigma(t-t_0)}, \tag{31}$$

which leads to the well-known Sabine formula. For comparison, the decay for the Green's function solution of Eq. (9) for an unbounded medium,[9]

$$G_u(\mathbf{r},\mathbf{r}_s,t) = \frac{P}{(4\pi D)^{3/2}} \frac{e^{-r^2/[4D(t-t_0)]}}{(t-t_0)^{3/2}} e^{-\sigma(t-t_0)}, \tag{32}$$

should be briefly investigated. A coarse analysis of this relation can distinguish two main behaviors. For smaller values of $(t-t_0)$, the decay is dominated by the term $e^{-r^2/[4D(t-t_0)]}/(t-t_0)^{3/2}$, which has some importance in the evaluation of the early decay time (EDT, i.e., the decay time based on the 0 to -10 dB part of the temporal sound-energy decay function). Meanwhile, the "*Sabine*" decrease $e^{-\sigma(t-t_0)}$ becomes predominant otherwise, and has more influence on the evaluation of the reverberation time (RT, based on the -5 to -35 dB part of the temporal sound-energy decay function). It is also noticeable that the dependence of the sound decay on r makes possible some spatial variations of the sound decay, which is an important matter for long or flat rooms, or for coupled rooms for example. This will be discussed further in the numerical results presented in the next section.

IV. NUMERICAL RESULTS AND IMPLEMENTATION

In Sec. IV A, the principle of the numerical implementation is presented for the case of a room with "proportionate" (similar) dimensions (here, a cubic room) and uniform absorption, for both steady-state response and sound decay. Results are compared with those given by the statistical theory, which is valid in this case. An additional application is given for the case of sound transmission through a partition, for the same cubic room. In Sec. IV B, two examples are given to illustrate the capability of the method, for room shapes that usually lead to non-uniform sound fields: the calculation of the distribution of steady-state sound level in a flat room, and the calculation of sound decay in a long room. Results are compared with simulations obtained by the ray-tracing method.

A. Rooms with "proportionate" dimensions

1. Steady-state sound field

Both of the systems of Eqs. (13) and (14) (i.e., Model 1), or Eqs. (21) and (22) (i.e., Model 2), are solved numerically for a cubic room of dimensions $10\times10\times10$ m^3 with homogeneous surface-absorption coefficient 0.1. Let us define an omnidirectional sound source located at the geometrical center of this room, radiating a constant sound-power level of 100 dB, i.e., a power output $W_s=10^{-2}$ W.

The numerical method used for solving the diffusion equation together with the boundary conditions is based on the finite element method (FEM). The software used to obtain the results presented in Sec. IV is *Femlab 3.1i*. First, a cubic domain representing the room is defined and meshed with about 5.10^3 unstructured, linear, Lagrange-type elements. A small spherical subdomain V_s of volume v and radius r, centered on the source location is defined. The subdomain V defined by the cubic domain minus V_s is then the domain in which the sound-energy density is to be calculated.

Let us mention here that the most important limitation of finite-element methods (i.e., the size of the elements) is not strictly a problem with this model, since the size of the elements is more dependent on the mean free path than on the wavelength—which determines the maximum size of the elements in the case of solving the Helmholtz equation. The same meshing can be applied for all frequency bands, since frequency is only taken into account in the absorption coefficients of the room boundaries, as already mentioned above. However, the size of the elements should be on the order of (or smaller than) one mean free path. Thus, the diffusion model can be applied to very large enclosures with a limited meshing.

To model the problem, we solve the stationary form of the partial differential equation (13) of Model 1 in both subdomains with two different source terms:

$$-D\nabla^2 w(\mathbf{r}) + \sigma w(\mathbf{r}) = 0 \quad \text{in } V, \tag{33}$$

$$-D\nabla^2 w(\mathbf{r}) + \sigma w(\mathbf{r}) = \frac{W_s}{v} \quad \text{in } V_s. \tag{34}$$

The source terms of this system of equations model a power source as defined in Eq. (10), in the case of a constant power output, ensuring that the power supply integrated over V_s is equal to W_s. The associated boundary conditions are homogeneous Neumann conditions [Eq. (14)].

Alternatively, the solution of Model 2 can be calculated in the same way. The system of Eqs. (33) and (34) is solved without the volume absorption term $\sigma w(\mathbf{r})$, and with mixed-type boundary conditions [as defined in Eq. (22)] instead of homogeneous Neumann conditions, with $\alpha=0.1$.

From the numerical solution for $w(\mathbf{r})$, one can compute the local sound-pressure level as[14]

$$L_p(\mathbf{r}) = 10\log\left(\frac{w(\mathbf{r})\rho c^2}{P_{\text{ref}}^2}\right), \tag{35}$$

with $P_{\text{ref}}=2\times10^{-5}$ Pa, and ρ the air density. The solutions of both Models 1 and 2, in terms of sound-pressure levels for the reverberant field, are presented in Fig. 2(a), together with the constant value given by statistical theory, as expressed in Eq. (28). The calculation time is a few seconds on a standard computer. The radius of the small spherical subdomain V_s for the source is 0.2 m. Analytically, the Green's function of the stationary diffusion equation has a singularity at the source location, as expressed in Eq. (30). The numerical model defined by Eqs. (33) and (34) reduces this singularity. If the radius of the spherical domain V_s decreases, the peak height at the source location increases, and eventually becomes infinite when the subdomain is infinitely small. As explained in Sec. III C 2, the singularity at the source location has no physical meaning and its effect on the results has to be minimized. At a distance from the source greater than 1.5 m, the diffusion models match the statistical-theory solution very well. Another important observation from

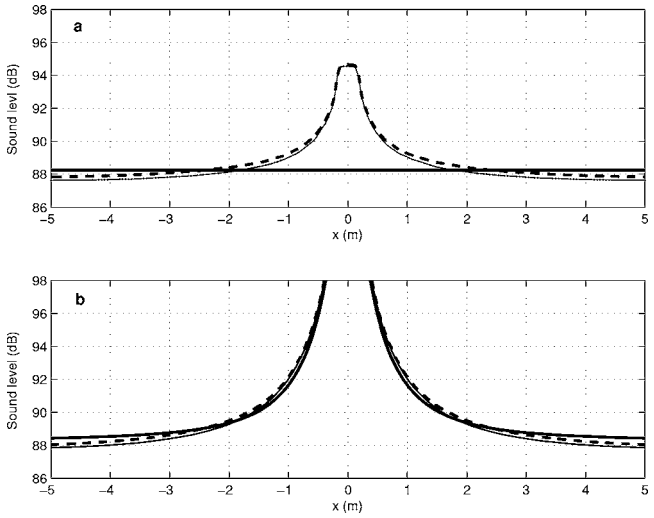


FIG. 2. Steady-state sound pressure level (dB) calculated from the numerical solution of the diffusion equation, for a $10\times10\times10$ m^3 cubic room ($\alpha=0.1$, source power level 100 dB), with model 1 (thin solid line), model 2 (dashed line), and the statistical theory (thick solid line): (a) reverberant field; (b) direct field plus reverberant field.

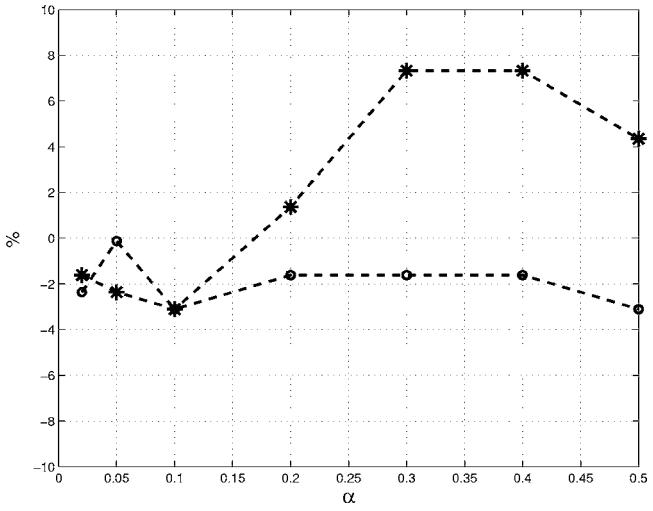


FIG. 3. Deviation in % of the reverberation times calculated by the diffusion model and the statistical theory (Sabine formula), for a $10\times10\times10$ m^3 cubic room, as a function of the wall absorption: (○) model 1; (∗) model 2.

Fig. 2(a) is that both Models 1 and 2 lead to almost identical solutions, as expected for a cubic room (see Sec. III B 2).

As shown in Fig. 2(b), the total sound-pressure solution, taking into account the direct-field sound contribution, is computed according to the relation:

$$L_p^{\mathrm{tot}}(\mathbf{r}) = 10\log\{\rho c[W_s/(4\pi r^2) + w(\mathbf{r})c]/P_{\mathrm{ref}}^2\}. \tag{36}$$

Both Models 1 and 2 are in very good agreement with diffuse sound-field theory. The direct field being dominant in the near field of the source, the peak of the diffusion solution does not corrupt the final solution. Practically, it is observed that a spherical subdomain V_s of radius of about 20 cm allows the removal of any effect of the peak of the diffusion solution at the source location [see Fig. 2(a)], after adding the direct field. This rule is true for all of the simulations that have been carried out with the diffusion model so far, and is applied to all further results presented in this study.

Some other simulations have shown that, even in the case of rooms with different aspect ratios (long or flat rooms), where the sound field is not uniform, both Models 1 and 2 lead to very similar results in terms of both steady-state levels and sound decay. This confirms that the boundary condition (22) of Model 2, which can be demonstrated in the case of a diffuse sound field,[2] is a good approximation for sound fields with substantial spatial variations. As mentioned earlier, the advantage of Model 2 is that the mixed boundary conditions relate the local energy loss at the boundary directly to the surface-absorption coefficient. It then allows rooms with walls having different absorption coefficients to be modeled (an example is given in Sec. IV B 1).

2. Sound-energy decay

To obtain the sound decay in a given room, Models 1 and 2 are solved numerically in their time-dependent forms. A simple way of computing the time decay is to set up different initial conditions in the subdomains V and V_s,

$$w(\mathbf{r},0) = 0 \quad \text{in } V, \tag{37}$$

$$w(\mathbf{r},0) = w_0 \quad \text{in } V_s, \tag{38}$$

where w_0 is the initial value.

The same cubic room as that studied in Sec. IV A 1 is investigated for a range of values of the surface absorption α. The room is discretized into about 10^3 elements (the calculation time is about 20 s). The numerical results give identical sound-decay functions at every point in the room. The obtained decay functions in dB are linear, in accordance with diffuse sound field theory, but start at $t=0$—a feature that will be discussed further in Sec. IV B 2. The reverberation time (RT), defined in Sec. III C 3, is evaluated from the sound decay function. The relative deviation from the RT given by Sabine's formula [i.e., $0.163V/(\alpha S)$] is presented in Fig. 3. Model 1 achieves deviations less than 4% for all absorption coefficients. The same deviation is verified for Model 2 only for coefficients of less than 0.25. For higher absorption it overestimates the RT by about 4–8% (note that practically all rooms have average absorption coefficients less than 0.2).

The numerical results have shown that the diffusion model matches well the statistical theory in the case of a cubic room. In Sec. IV B, some results for disproportionate rooms, where the acoustic energy and sound decay are no longer uniform in the room (i.e. the statistical theory is not valid), will be considered.

3. Sound transmission through a partition

Let us consider the case of two adjacent rooms separated by a partition with transmission loss TL (in dB), defined as follows:

$$TL = 10\log\left(\frac{1}{\tau}\right), \tag{39}$$

where τ is the partition transmission coefficient. A source is located in one of the rooms, and generates a diffuse sound field of energy density w_1. The local ingoing flux, as defined

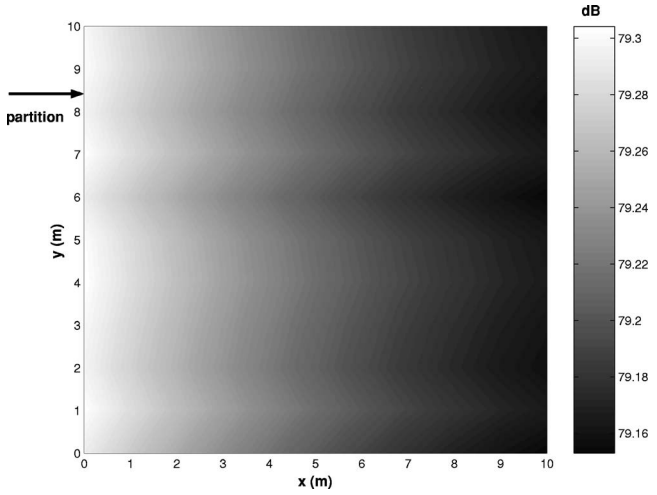


FIG. 4. Sound pressure-level distribution in a $10\times10\times10$ m^3 cubic room with uniform wall absorption 0.02. Sound is transmitted through a wall of transmission loss 30 dB at $x=0$ (left side).

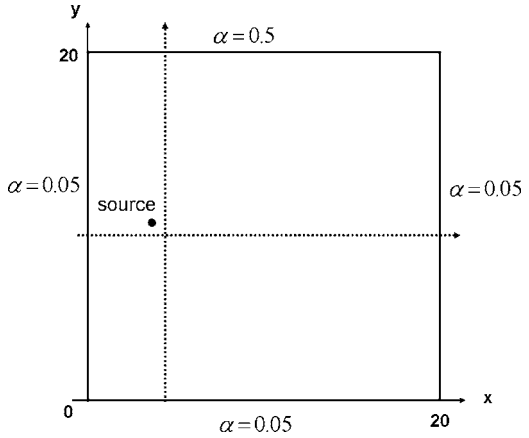


FIG. 5. Schematic of the $20\times20\times2.5$ m^3 flat-room case study (upper view). Pressure level is plotted along the two dotted lines at $y=9.5$ m and $x=4.5$ m.

in Eq. (12), is then $J_{\text{in}}=\tau cw_1/4$.[2] The boundary condition for the partition must take into account both sound-absorption phenomena [as in Eq. (22) for Model 2] and sound transmission [Eq. (12)], and may then be written

$$D\frac{\partial w(\mathbf{r},t)}{\partial n}+\frac{c\alpha}{4}w(\mathbf{r},t)-\tau cw_1/4=0. \tag{40}$$

The former example of a cubic room of dimensions $10\times10\times10$ m^3 with a uniform absorption coefficient of 0.02 is taken. One of the surfaces of the room connects to a room in which a source radiates a diffuse sound field of pressure level $L_{p1}=100$ dB. The partition (located at $x=0$) has a transmission loss $TL=30$ dB. Figure 4 presents the distribution of the sound-pressure level in the cubic room in a horizontal plane at height 5 m, using Model 2. As expected, the sound-pressure level in such a cubic room is rather constant (at about 79.2 dB), with a very slight decrease at increasing distance from the partition. The sound pressure L_{p2} in the room, as calculated from the statistical theory using the relation[1]

$$L_{p2}=L_{p1}-R-10\log(A/S_p), \tag{41}$$

gives a similar level of 79.2 dB, A and S_p being, respectively, the absorption area of the cubic room and the area of the partition. This method can easily be extended to an elongated room shape, and could treat the case of sound transmission into a hall through windows. This practical case is often encountered when a residential building is located close to a major road. Let us also mention that the method can accommodate as many partitions with sound transmission, and as many point sources, as necessary to model any given practical case.

B. Examples of long and flat rooms

In the following, two applications for "disproportionate" rooms (rooms with dissimilar dimensions) are presented: the calculation of the steady-state sound field in a flat room, and the calculation of the sound-decay variations in a long room. Numerical solutions for the diffusion equation are presented for Model 2. The calculation times are about 5 s and 1 min for obtaining, respectively, the value of the sound-pressure level (steady-state calculation) and sound decay at any location in the rooms studied. As the spatial variations of the sound field are substantial in these cases, the statistical theory is no longer applicable. For such disproportionate rooms, the proposed method has been validated by numerous comparisons with reference results (two examples are presented in this paper). For doing such comparisons, it is convenient to use some simulation results obtained from a software that has been validated in the literature.[11,15] Future studies will focus on an experimental validation of the method, in cases such as long rooms and coupled rooms.

For comparison, the RAYSCAT software[15] based on the ray-tracing theory is used. The diffusion theory uses the classical mean free path which is based on purely diffuse surface reflections (see Sec. II). For this reason, in the *Rayscat* predictions, reflections are modeled as purely diffuse (i.e., governed by Lambert's law). In the steady-state simulations the number of rays is 10^4, and the maximum number of reflections per ray[16] is 20 (calculation time 10 s for a line of receptors). For the sound-decay calculations the number of rays, and the number of reflections per ray, were set to 2×10^4 and 200, respectively (calculation time about 1 min for one receptor). Results achieved good convergence with this choice of parameters.

1. Sound attenuation in a flat room

The case which is considered here is a flat room of dimensions $20\times20\times2.5$ m^3 (Fig. 5). The walls, ceiling and floor have a low absorption coefficient of 0.05, except for the wall at $y=20$ m, which is highly absorbent ($\alpha=0.5$). Remember here that such a case of a room with surfaces having different absorption coefficients can only be treated with Model 2. A point source of constant output power level 100 dB is located at point (4,10,1). The sound-pressure level is investigated at a height of 1 m along two lines passing close to the source—at $y=9.5$ m and $x=4.5$ m, as shown in Fig. 5.

The agreement between the diffusion model and the ray-tracing model is good, with a maximum discrepancy of 1 dB

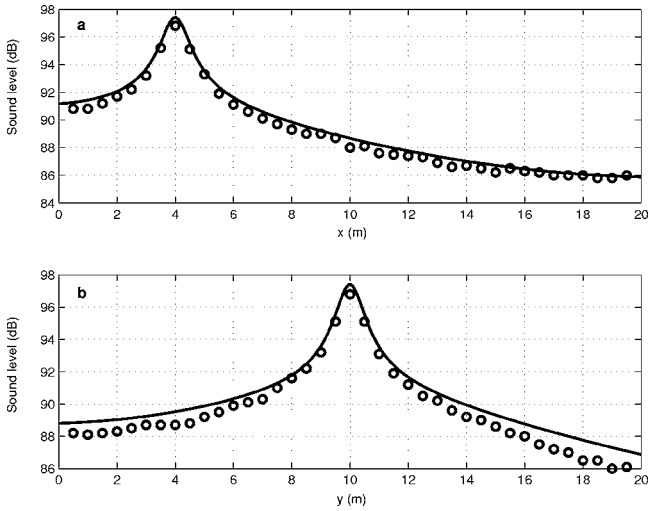


FIG. 6. Sound pressure level (dB) in a flat room $20\times20\times2.5$ m^3 (see Fig. 5 for details about geometry and absorption). Sound level is plotted at height 1 m, along two lines $y=9.5$ m (a) and $x=4.5$ m (b), with the diffusion model (solid line) and a ray-tracing method (○).

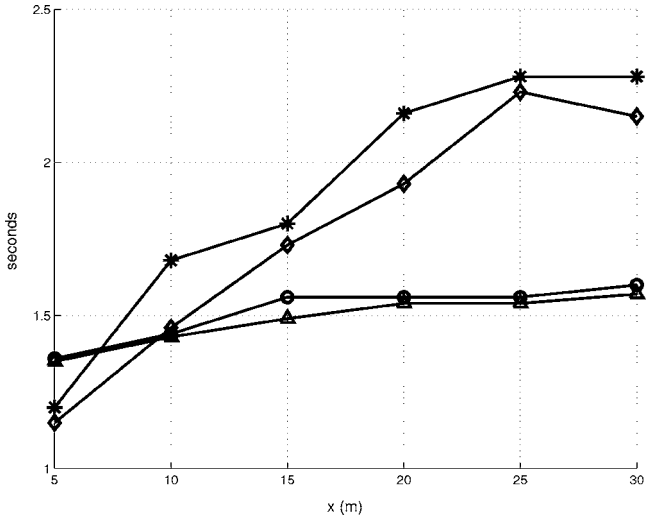


FIG. 7. RT and EDT calculated in a long room $4\times4\times32$ m^3 with uniform absorption coefficient 0.1, as a function of the receiver position along the longer dimension of the room. The source is located at $x=2$ m. (○) RT diffusion; (△) RT ray-tracing; (∗) EDT diffusion; (◇) EDT ray-tracing.

(see Fig. 6). In the x direction [Fig. 6(a)], both models indicate the same significant spatial attenuation of the pressure level. In the y direction [Fig. 6(b)], both models show a higher sound attenuation (by about 2 dB) close to the absorbing wall (located at $y=20$ m). The attenuation given by the ray-tracing model is about 1 dB higher than the diffusion model, but the general agreement is good. Additional results (not presented here) showed that, for flat or long rooms with mean absorption coefficient higher than 0.2, the sound attenuation in dB becomes substantially higher for the diffusion model than for the ray-tracing model. However the range [0,0.2] for the mean absorption coefficient covers the majority of cases in practical applications.

2. Reverberation time in a long room

It is well-known that, in long rooms, the temporal sound decay depends on the distance of the receiver to the source.[3] This phenomenon is addressed in the present example. A hall of dimensions $4\times4\times32$ m^3 with uniform absorption coefficient 0.1 is studied. The source is located at $x=2$ m (i.e., 2 m from one of the extremities of the room), x being the axis along the longest dimension of the hall, from 0 to 32 m. The sound decay is investigated at positions $x=5$, 10, 15, 20, 25, and 30 m, in terms of both RT and EDT (as defined in Sec. III C 3). The time-dependent form of the diffusion equation (Model 2) is solved, as explained in Sec. IV A 2.

Figure 7 shows the RT and EDT estimated with both diffusion and ray-tracing models. The reverberation times given by both models are in very good agreement, and show a slight increase of about 15% along the length of the room. Meanwhile, the EDT given by the diffusion models tends to be 15% higher than the one given by the ray-tracing model. In any case, the trend for both models is the same, and indicates a much higher increase for the EDT than for the RT, reaching about 100% along the room length. That is, the early part of the sound decay is more affected by the room shape.

An investigation of the time-solution curves given by the diffusion model for positions $x=5$, 15, and 25 m (Fig. 8), together with the investigation of the Green's function in Eq. (32), gives further insight into this result. Near the source ($x=5$ m) the decay is mostly governed by the decrease $e^{-\sigma t}$. Further from the source, the t-increasing function $e^{-r^2/4Dt}$ is more and more apparent for the early part, due to the increase of r (source-receiver distance). Physically, the increasing first part of the time-solution models the dispersion of the arrival times of the first sound particles reaching the receiver. This dispersion is higher when the receiver is further from the source. The increasing first part of the time-solution makes the early part of the decrease (i.e., the early part of the sound decay) flatter, and then induces a very significant rise in the EDT. The later part of the decay is, for all positions, governed by the same decrease in $e^{-\sigma t}$. The $1/t^{3/2}$ term in the

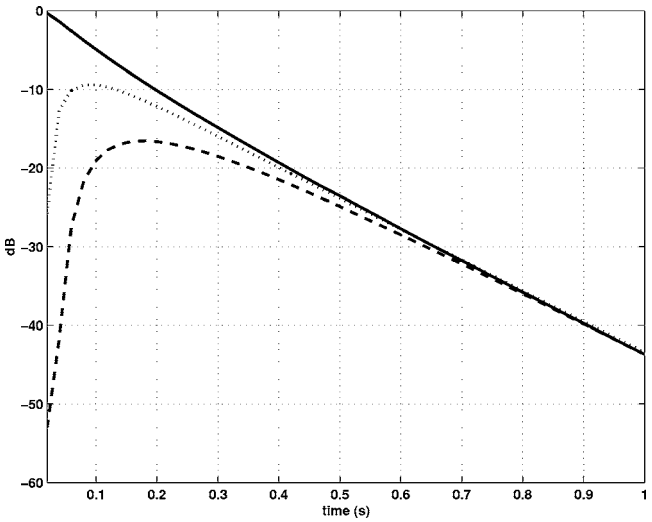


FIG. 8. Sound decay curves evaluated by the diffusion model in a $4\times4\times32$ m^3 hall (uniform absorption 0.1, source located at at $x=2$ m) at several locations. $x=5$ m (solid line), $x=15$ m (dotted line), and $x=25$ m (dashed line).

Green's function [Eq. (32)] seems to have little influence on the main features of the observed time-decay curves: this term could be responsible for a steep decay in the early part of the sound-decay function, but such feature is not visible in the obtained curves.

It is also noticeable that the sound decay functions of Fig. 8 start at $t=0$, which is obviously nonphysical, as the sound-decay function should start after the arrival of the direct sound. This feature can be explained by the nature of the Green's function of Eq. (32): this function indicates that the source has instantaneous effects at any location, but, as pointed out by Morse and Feshbach,[9] it is not valid theoretically for times smaller than the mean free time, i.e., λ/c. The early part of the obtained time-decay functions should then be considered with caution. However, the RT and EDT are calculated in this paper from the maxima of the sound-decay functions (for example from $t=0$ for $x=5$ m in Fig. 8, and from $t\simeq 0.9$ s for $x=15$ m), and are in good agreement with ray-tracing data.

V. CONCLUSION

In the initial theoretical part of this paper (Sec. III), the use of a diffusion process to model the reverberant sound field in empty rooms has been discussed. In addition to recent work dealing with an analogy between the diffusion of particles by a medium of scatterers and room acoustics,[4] the paper proposes a generalization of the model to arbitrary room shapes. It focuses on the definition of appropriate source terms of the diffusion equation for room acoustics, and on a choice of ad hoc boundary conditions. It is then shown that the proposed model may be considered as a natural extension of the classical statistical theory of reverberation.

In the second part (Sec. IV), the numerical application of the diffusion model is discussed in the context of room-acoustical applications, and is compared with the statistical theory, in the case of a cubic room. The numerical results show a good agreement with statistical theory, both in terms of sound-pressure levels and sound decay. Two additional examples are then given, involving the prediction of steady-state pressure levels in a flat room and the spatial variation of the sound decay in a long room. Here again, the results are in good agreement with numerical results obtained by a ray-tracing model.

The main interest of this approach is its ability to give satisfactory results whatever the room shape, at any receiver location, with a calculation time on the order of seconds only for the steady state, and on the order of minutes for a time-varying state, while most existing software needs much longer computational times. Moreover, the most important limitation of finite-element methods (i.e., the size of the elements) is not strictly a problem in this case, since the size is more dependent on the mean free path than on the acoustic wavelength. Thus, the diffusion model can be applied to very large enclosures with a limited meshing. Such an approach could be used in the first phase of an architectural project, to calculate and design the main acoustical features of a room (sound-field distribution and decay), while a classical approach, such as ray-tracing, could be used in the second phase, to give more detailed and accurate results.

ACKNOWLEDGMENT

The authors would like to thank *Région Poitou-Charentes* for sponsoring the visit of V. Valeau to the University of British Columbia from February to August 2005.

[1] L. Cremer and H. Muller, *Principles and Application of Room Acoustics* (Applied Science Publishers, London, 1978), Vol. 1.
[2] H. Kuttruff, *Room Acoustics*, 4th ed. (Spon Press, London, 1999).
[3] J. Kang, *Acoustics in Long Spaces: Theory and Design Guide* (Thomas Thelford Publishing, London, 2002).
[4] J. Picaut, L. Simon, and J.-D. Polack, "A mathematical model of diffuse sound field based on a diffusion equation," Acust. Acta Acust. 83, 614–621 (1997).
[5] J. Picaut, L. Simon, and J.-D. Polack, "Sound field in long rooms with diffusely reflecting boundaries," Appl. Acoust. **56**, 217–240 (1999).
[6] J. Picaut, J. Hardy, and L. Simon, "Sound field modeling in streets with a diffusion equation," J. Acoust. Soc. Am. **106**, 2638-2645 (1999).
[7] W. Joyce, "Sabine reverberation time and ergodic auditoriums," J. Acoust. Soc. Am. **58**, 643-655 (1975).
[8] F. Ollendorf, "Statistische raumakustik als diffusionsproblem," Acustica **21**, 236 (1969).
[9] P. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, 1953).
[10] F. Hunt, "Remarks on the mean free path problem," J. Acoust. Soc. Am. **36**, 556–564 (1964).
[11] M. Hodgson, "Evidence of diffuse surface reflections in rooms," J. Acoust. Soc. Am. **89**, 765–771 (1991).
[12] E. Gilbert, "Reverberation in spherical rooms with nonuniform absorption," Acustica **66**, 275-280 (1988).
[13] Y. Adda and J. Philibert, *La Diffusion dans les Solides* (Presses Universitaires de France, Paris VI, 1966), Vol. 1.
[14] A. D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications* (Acoustical Society of America, New York, 1981).
[15] A. Ondet and J. Barbry, "Modeling the sound propagation in fitted workshps using ray tracing," J. Acoust. Soc. Am. **85**, 787–796 (1989).
[16] M. Hodgson, *(D,E)Raycub User's Manual*.

Imaging nonlinear scatterers applying the time reversal mirror

T. J. Ulrich[a] and P. A. Johnson[b]
Los Alamos National Laboratory of the University of California, EES-11, Los Alamos, New Mexico 87545

A. Sutin[c]
Stevens Institute of Technology, Artann Laboratories

(Received 12 August 2005; revised 22 December 2005; accepted 1 January 2006)

Nonlinear elastic wave spectroscopy (NEWS) has been shown to exhibit a high degree of sensitivity to both distributed and isolated nonlinear scatterers in solids. In the case of an isolated nonlinear scatterer such as a crack, by combining the elastic energy localization of the time reversal mirror with NEWS, it is shown here that one can isolate surfacial nonlinear scatterers in solids. The experiments presented here are conducted in a doped glass block applying two different fixed frequency time-reversed signals at each focal point and scanning over a localized nonlinear scatterer (a complex crack). The results show a distinct increase in nonlinear response, via intermodulation distortion, over the damaged area. The techniques described herein provide the means to discriminate between linear and nonlinear scatterers, and thus to ultimately image and characterize damaged regions. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2168413]

PACS number(s): 43.60.Tj, 43.25.Dc, 43.40.Le, 43.40.Fz [MFH] Pages: 1514–1518

I. INTRODUCTION

Combining the time reversal mirror (TRM) with elastic nonlinearity holds great promise for isolating a nonlinear scatterer such as a crack in a solid. The TRM provides the means to narrowly focus wave energy in time and space. If the focal point is a nonlinear scatter, harmonics and other nonlinear effects are induced that are not induced elsewhere in an otherwise elastically linear solid. Thus the combination of illuminating nonlinear scatterers applying time reversal (TR) and analyzing the focused signal for nonlinear response offers the means to image them. Other wave scatterers such as sidewalls and voids do not exhibit nonlinear response.[1] In this paper we show how the method can be applied to provide a high-resolution image of a crack located on the surface of a glass block. In the following, we present pertinent background information and then describe the experiment. Following this, we present the results and discussion and then conclude.

II. BACKGROUND

Time reversal has been taken to an advanced state primarily by the group at the University of Paris VII (Laboratoire Ondes et Acoustique, ESPCI) who have comprehensively explored the underlying fundamental physics and have developed numerous applications.[2–6] TRM provides the means to focus an ultrasonic wave, regardless of the position of the initial source and regardless of the heterogeneity of the medium in which the wave propagates. TRM has a wide range of applications, including destruction of tumors and kidney stones and long-distance communication in the ocean.[7] The nondestructive evaluation (NDE) applications of (linear) TRM to date include detection of small, low-contrast defects within titanium alloys[4,5] and detection of cracks in a thin air-filled hollow cylinder.[6] In these experiments nonlinear effects were not considered.

Over the last two decades, studies of nonlinear wave methods applied to nondestructive evaluation (NDE) have demonstrated that the nonlinear elastic response of materials may radically increase in the presence of damage (cracks or other flaws).[1,8–11] Numerous applications to NDE have been developed by many groups in the US, Europe, and Russia.

TRA focusing, providing a large localized amplitude of an acoustic wave, is a good tool for investigation of nonlinear effects for the application for NDE. In the work of Montaldo *et al.*[12] the second harmonic generation in the TRA focused signal in homogeneous fluid media was observed. Here we employ nonlinear interaction of the TRA focused signal with a crack in a solid (i.e., a highly nonlinear scatterer) for crack detection purposes applying the wave modulation technique. To our knowledge the experiment described here and its predecessors[13,14] are the first to implement nonlinear time reversal and apply it to NDE and imaging in solids.

III. EXPERIMENTAL DETAILS

One of the most straightforward methods to evaluate the nonlinear elastic response of a material is to measure the modulation of an ultrasonic wave (f_{hi}) by a low-frequency vibration (f_{lo}). This gives rise to nonlinear wave mixing in materials with distributed (e.g., a rock) or localized damage (e.g., a crack in a solid) where the spectral composition contains both pure tones (f_{lo} and f_{hi}) as well as sum and difference frequencies ($f_+ = f_{lo} + f_{hi}$ and $f_- = f_{hi} - f_{lo}$), while an undamaged (elastically linear) material would simply exhibit the pure tones. These additional frequencies in the signal (f_+

[a]Electronic address: tju@lan1.gov
[b]Electronic address: paj@lan1.gov
[c]Electronic address: asutin@stevens-tech.edu

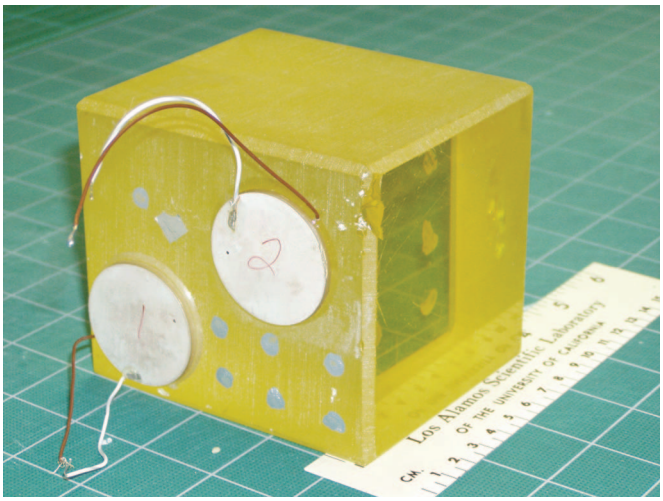


FIG. 1. Doped glass block showing the attached sources. The sources are located approximately symmetrically about the crack on the opposing face. A cm scale is located on the right of the sample.

and f_-) are usually referred to as sidebands of the high frequency because $f_{lo} \ll f_{hi}$, and thus the sum and difference frequencies lie equally and closely spaced about f_{hi}. In principle, however, the only requirement is that f_{lo} and f_{hi} are different enough to resolve, thus f_+ and f_- do not necessarily lie close to f_{hi} as is typical of the concept of sidebands. This is true of the frequencies used in the experiment presented here ($f_{lo}=170$ kHz, $f_{hi}=255$ kHz, $f_+=425$ kHz, and $f_-=85$ kHz). These frequencies were chosen from the sample's natural frequencies by selecting the frequencies corresponding to the highest amplitudes in response to a broadband pulse.

The experiment was conducted in a doped glass block of dimensions $101\times89\times89$ mm ($\rho\approx3.0$ g/cm^3, $c\approx2.5$ km/s, $Q\approx2000$). Two piezoelectric ceramics (PZT-5a, 38 mm diameter, 2.8 mm thickness) were used as sources. The sources were bonded to the face of the glass block opposite to the crack location. Photographs of the sample are found in Figs. 1 and 2, showing the cracked face and attached ceramics. The elastic response on the cracked face was measured using a broadband (DC to 1.5 MHz) laser vibrometer (Polytec model OFV 303, controller OFV 3001, velocity range 1 V=1 m/s). A block diagram of the apparatus is shown in Fig. 3.

Each source was driven with a different frequency [$f_1=f_{lo}=170$ kHz and $f_2=f_{hi}=255$ kHz, signal lengths $\tau=0.1$ ms, sinusoidal envelope, see Fig. 4(a)]. The direct signal was measured for each source independently and then time reversed. This experiment differs from standard TR experiments as the direct time-reversed signals were then input back at the original source locations, as was done in other work.[15–17] Reciprocity dictates that the resulting signal will focus at the point of the original detector.[18] This fact allows for the use of multiple low-amplitude channels to produce a focused (in space and time) high-amplitude response at any point in the solid. In these experiments we restrict ourselves to focusing on the surface of the sample; as such, the crack analyzed here is a surface crack with a penetration depth of a few millimeters.

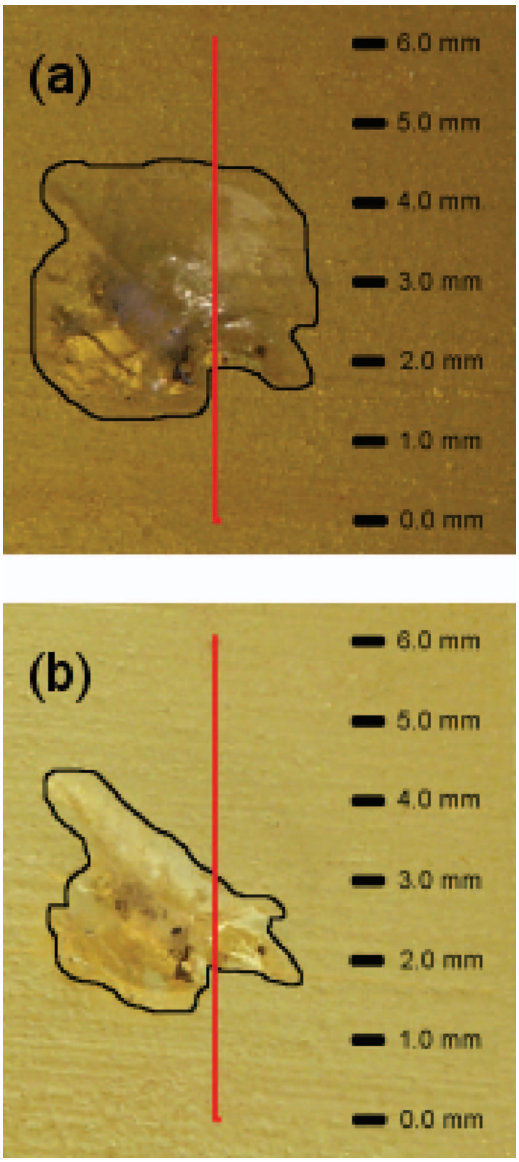


FIG. 2. Close-up of the crack in the doped glass block (a) lit from the surface and (b) backlit through the glass. The vertical red line indicates the path along which the measurements were made. The visually apparent extent of the crack is denoted by the black outline encircling the cracked region. The two lighting techniques illuminate different features: the surface lighting illuminates the two-dimensional surfacial features while the backlighting highlights the third dimension of depth but eliminates some of the surface detail. Notice the difference in apparent crack extent with the change of lighting as well as the obvious complexity of the crack (e.g., crack shape and density variations).

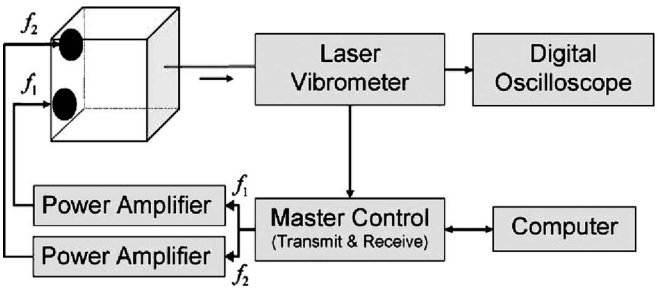


FIG. 3. Block diagram of the nonlinear TRA (NTRA) experimental system.

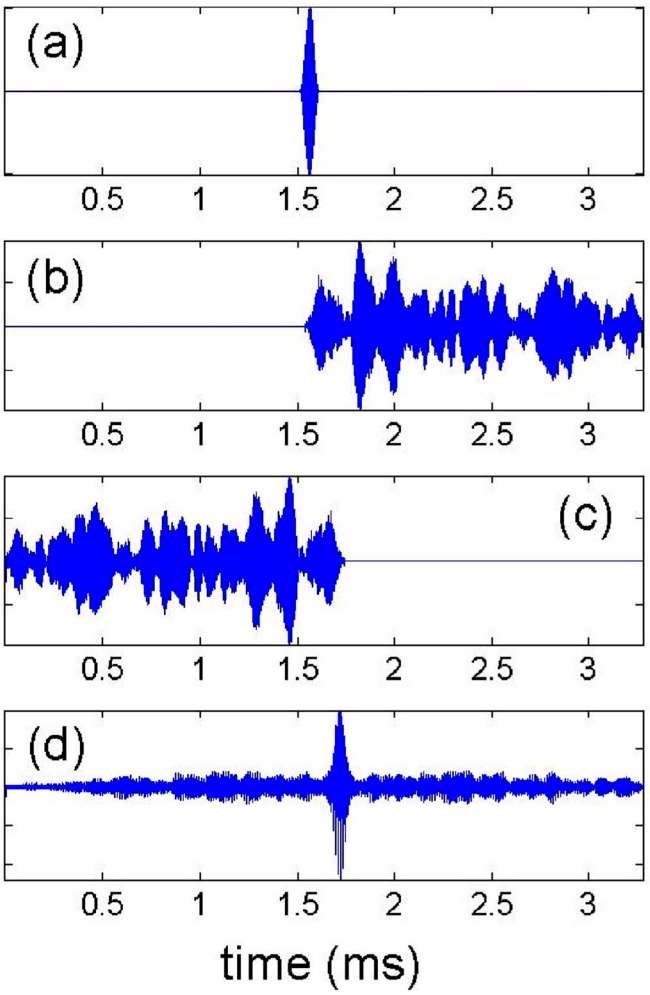


FIG. 4. (Color online) Waveforms from the TR experiment: (a) Input signal $s_n(t)$ for f_1 ($n=1$, f_2 not shown); (b) signal received at the point of detection $r_{m,n}(t)$ (from the f_1 source); (c) TR signal $r_{m,n}(-t)$ input to the transducer at the location of the original source f_1; and (d) focused signal R_m at the original point of detection due to the transmission of the TR signals [i.e., (c) and the analogous signal from f_2, not shown].

The experimental protocol of the nonlinear TR method used in this experiment is as follows:

(1) Transmit signal $s_n(t)$ from source n, Fig. 4(a),

$$s_n(t) = U_n \cos(2\pi f_n t) \sin^2\left(\frac{\pi t}{\tau}\right), \tag{1}$$

where f_n is the frequency transmitted from the nth transducer with a maximum amplitude U_n and duration τ.

(2) Receive the signal $r_{m,n}(t)$ at the selected point x_m (i.e., focal point of the laser vibrometer) on the sample for each f_n independently, e.g., Fig. 4(b).

(3) Time reverse the received signals, Fig. 4(c),

$$r_{m,n}(t) \rightarrow r_{m,n}(-t). \tag{2}$$

(4) Simultaneously transmit the amplified TR signals $S_{m,n}(t)$ from their original source locations,

$$S_{m,n}(t) = U_n \frac{r_{m,n}(-t)}{\max(r_{m,n}(-t))}, \tag{3}$$

where the U_n's are the same maximum amplitudes used in step 1.

(5) Record the TR focused signal $R_m(t)$ at the mth focal point of the laser vibrometer (x_m), Fig. 4(d),

$$R_m(t) = \sum_{n=1}^{N} \int_0^{t_{\text{final}}} dt' G(x_m t | x_n t') S_{m,n}(t'). \tag{4}$$

N is the total number of source transducers, x_n is the location of the nth transducer, and $G(x_m t | x_n t')$ is the Green's function taking the signal transmitted from location x_n and time t' to point x_m at a later time t. The integration is taken over the entire length of the waveform from 0 ms to t_{final} = 3.2768 ms (i.e., 32 768 samples taken at 10 MS/s).

(6) Move the laser focal point to the next desired location (x_{m+1}) and repeat the process for M total focal points.

The above steps are the generalized procedure followed for the scan presented in this paper. In this experiment a scan (in one dimension) was conducted across the cracked area. At each of the $M=22$ points (x_m) the same two (i.e., $N=2$) f_1 and f_2 sources were used. Additionally the drive amplitudes and durations remained constant at $U_1=40$ V, $U_2=100$ V, and $\tau=100$ μs, as did the source locations x_1 and x_2, as the transducers were bonded to the surface of the sample. The difference in the drive amplitudes U_1 and U_2 arises due to the use of two different amplifiers with differing gain. The focal point resolution (i.e., spot size of the laser) is ~ 10 μm. Scanning across the crack was done with graduated increments. In and near the cracked region the step size was consistently 200 μm, while away from the crack the step size varied from 200 to 500 μm. The path of the one-dimensional scan can be seen in Fig. 2.

IV. RESULTS AND DISCUSSION

The waveforms of the focused signals were saved at each focal point and analyzed for the presence of nonlinear scattering as well as other characteristic features that may be present. This was done by first comparing the complete set of unfiltered waveforms as a function of position [Fig. 5(a)]. Subsequently, these waveforms were then filtered about the frequencies corresponding to the sum and difference frequencies and again compared as a function of position [Figs. 5(b) and 5(c)]. The cracked region was visually measured to be in the range of $\sim$1.9–4.4 mm along the scan path (Fig. 2). The amplitudes of the focused waveform are found to increase near the region of the crack as do the sidelobe amplitudes. By comparison, the nonlinear features exhibit narrower focusing, indicating the extent of the crack boundaries (Fig. 5). In addition, the sum and difference frequencies display different behavior inside the crack. As the signal is focused at different locations along the crack, the difference frequency waveforms become more structured in amplitude than the sum frequency waveforms and exhibit more sidelobes [Fig. 5(b), positions 1.9–4.4 mm]. While this is still under investigation, it may relay information about the internal structure of the crack.

A spectral analysis of the waveforms was performed at each position to investigate the possibility of other frequencies being present in the signal (Fig. 6). The two primary frequencies are clearly observed through the entire scan and the sum and difference frequencies are seen inside the crack.

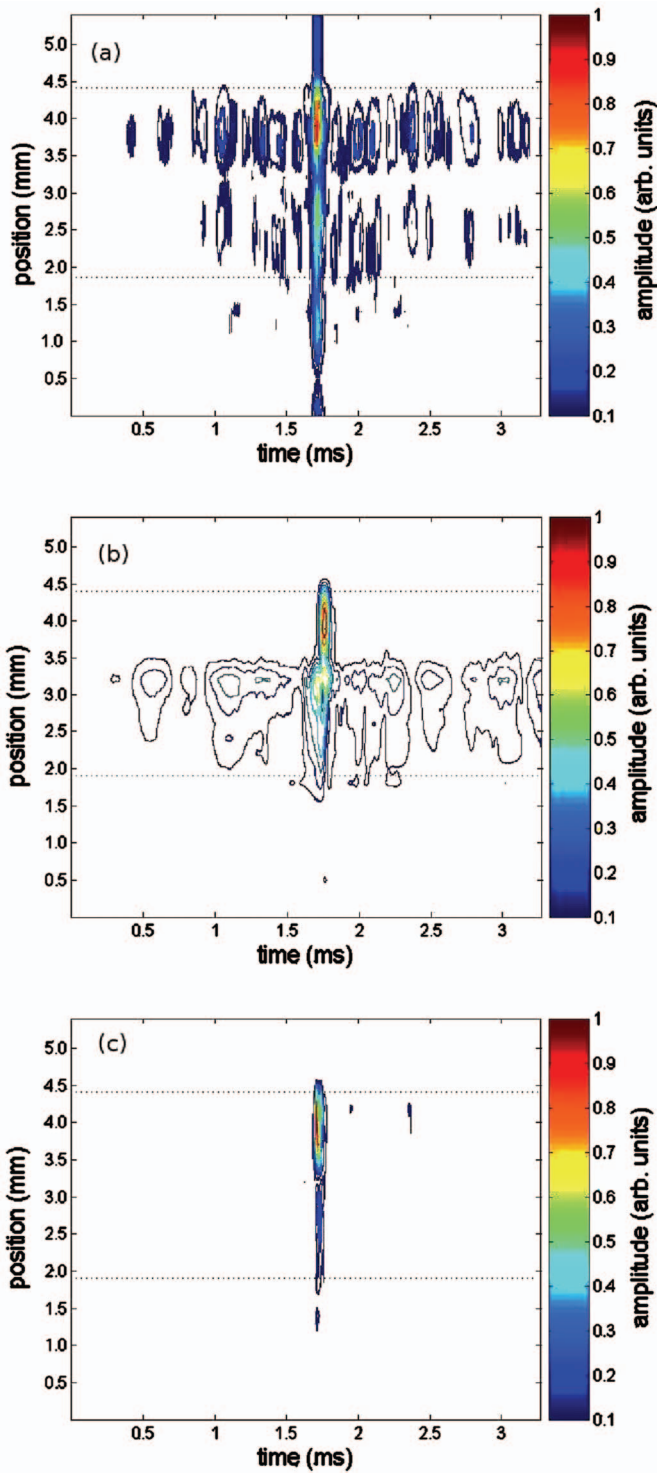


FIG. 5. Focused signal waveforms R_m at each focal position. The amplitudes of the waveforms increase in the damaged region (denoted by horizontal dotted lines), as well as gain more structure in the focused signal (i.e., the residual temporal lobes). (a) Unfiltered waveform; (b) waveform filtered about the difference frequency (f_-=85 kHz) with a bandwidth of 8.5 kHz (i.e., 10% of the filtered frequency); and (c) waveform filtered about the sum frequency f_+=425 kHz with a bandwidth of 42.5 kHz (i.e., 10% of the filtered frequency).

This plot also shows energy at 510 kHz, which is both $3f_1$ and $2f_2$. The image also shows complexity in broadening of energy around especially f_2 at two locations along the scan inside the cracked region for reasons that may have to do with crack complexity.

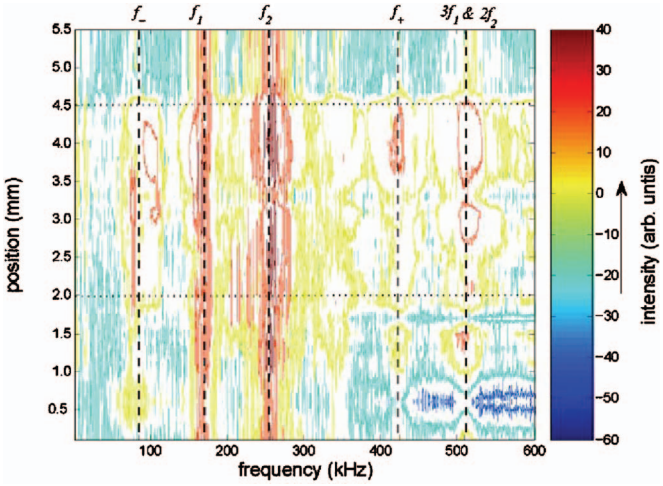


FIG. 6. Frequency components in the focused signal at each focal position. Notice the addition of sum ($f_+=f_1+f_2$=425 kHz) and difference ($f_-=f_2-f_1$=85 kHz) frequencies and harmonics (e.g., 510 kHz) in the damaged region. Frequencies of note are indicated by a vertical dashed line. The approximate extent of the damaged region is depicted by the horizontal dotted lines.

Three prominent features are of note in the focused waveforms. (1) The amplitudes of the waveforms in and around the crack are markedly higher than at points away from the crack (Fig. 5). (2) Waveforms in the cracked region contain the nonlinear signature of sum and difference frequency spectral components [Figs. 5(b), 5(c), and 6]. (3) In portions of the crack the TR focused signal is highly structured in its amplitude, i.e., large sidelobe amplitudes relative to the primary focused signal amplitude [Figs. 5(a), 5(b), and 7].

The additional structure of the residual temporal lobes (or sidelobes) in the TR focused waveforms is not simply a signal-to-noise issue where the sidelobes become more visible due to the increased amplitudes of the signals in the damaged area, rather, the amount of energy in these lobes E_{lobes} (in comparison to the total energy E_{total}) is much greater in the damaged/nonlinear region ($E_{\text{lobes}}/E_{\text{total}} \approx 0.88$) than in the undamaged material as ($E_{\text{lobes}}/E_{\text{total}} \approx 0.36$). The energy ratios at each point m were calculated as

$$\left.\frac{E_{\text{lobes}}}{E_{\text{total}}}\right|_m = 1 - \frac{\Sigma_{j=\alpha}^{\beta} R_m^2(t_j)}{\Sigma_{i=1}^{N_S} R_m^2(t_i)}, \tag{5}$$

where N_S is the number of samples in the waveform; α and β are the indices corresponding to the width of the initial pulse centered around the time of focus (t_{focus}=1.55 ms) thus defining the red portions of Fig. 7. All other parameters are as previously defined. The reason for the bleeding of the total energy into the sidelobes is unknown. One speculation is that this is further confirmation of nonlinearity of the signal through a nonlinear attenuation mechanism, such as that seen in rocks.[19] One nonlinear effect connected with interaction of the high-amplitude TRA focused wave with a crack is amplitude-dependent attenuation. This nonlinear attenuation leads to the decrease of the relative levels of the primary focused signal (Fig. 7, red portion) to the sidelobe amplitudes (Fig. 7, blue por-

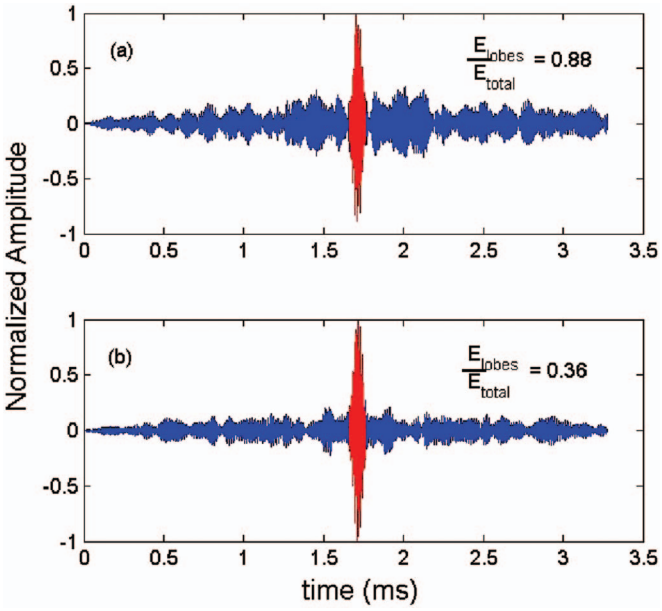


FIG. 7. TR focused waveforms (a) inside the crack and (b) outside the crack. The temporal sidelobes (blue) of the focused signal contain more energy in the cracked region (a) than outside of the crack (b). The width of the original signal [i.e., $s_n(t)$] is shown in red. In a perfect TR focusing the original source(s) would be reconstructed in time and space, thus the red portion of the plots (the primary focusing) is chosen to be the width of the original signal. The ratio of the energy in the temporal lobes (E_{lobes}) to the total energy of the signal (E_{total}) is given for each signal in the corresponding frame.

tion) of the TRA signal in the cracked area. This amplitude could be less than relative amplitude of the TRA signal out of the crack where there is no nonlinear attenuation.

While the presence of nonlinear components of the waveform in the damaged area did not come as a surprise, we were, however, surprised at the degree of amplification of the entire focused signal on the nonlinear scatter. We presume this is due to amplification due to the cracked region having the ability to oscillate freely, independent of bulk response. It appears that this effect is the result of the cracked surface having a tenuous connection to the block, thus focusing the same amount of acoustic energy upon the crack as was focused at an undamaged location, which will result in a greater amplitude, as we see here. The general amplification of the overall signature may then be a surface effect and one not expected in a nonlinear scatterer located internally; however, the presence of wave mixing and harmonics should be present.

We are continuing investigations into the effects of sample geometry, transducer locations, source amplitude effects, as well as the further study of the crack features presented in this paper. The development of a two-dimensional automated scanning system for full surface imaging is also underway to image surface cracks with a resolution of $\sim 10\ \mu$m. Additional efforts are being directed into probing subsurface damage features as well as intrinsic nonlinear properties in three-dimensional elastic solids.

ACKNOWLEDGMENTS

We are grateful to James TenCate and Robert Guyer. This research was supported in part by institutional support (LDRD) at the Los Alamos National Laboratory and by the United States Air Force Office of Scientific Research.

[1]P. A. Johnson, "The new wave in acoustic testing," Mater. World **7**, 544–546 (1999).
[2]M. Fink, D. Cassereau, A. Derode, C. Prada, P. Roux, M. Tanter, J. L. Thomas, and F. Wu, "Time-reversed acoustics," Rep. Prog. Phys. **63**, 1933–1995 (2000).
[3]M. Fink, "Time reversed acoustics," Sci. Am. **281**, 91–113 (1999).
[4]C. Prada, E. Kerbat, D. Cassereau, and M. Fink, "Time reversal techniques in ultrasonic nondestructive testing of scattering media," Inverse Probl. **18**, 1761–1773 (2002).
[5]N. Chakroun, M. Fink, and F. Wu, "Time reversal processing in nondestructive testing," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **42**, 1087–1098 (1995).
[6]E. Kerbrat, D. Clorennec, C. Prada, D. Royer, D. Cassereau, and M. Fink, "Detection of cracks in a thin air-filled hollow cylinder by application of the d.o.r.t. method to elastic components of the echo," Ultrasoni. Int. **40**, 715–720 (2002).
[7]W. A. Kuperman, W. S. Hodgkiss, H. C. Song, T. Akal, C. Ferla, and D. R. Jackson, "Phase conjugation in the ocean: Experimental demonstration of an acoustic time-reversal mirror," J. Acoust. Soc. Am. **103**, 25–40 (1998).
[8]O. Buck, W. Morris, and J. Richardson, "Acoustic harmonic generation at unbonded interfaces and fatigue cracks," Appl. Phys. Lett. **33**(5), 371–373 (1978).
[9]J. Cantrell and W. Yost, "Acoustic harmonic-generation from fatigue-induced dislocation dipoles," Philos. Mag. A **69**(2), 315–326 (1994).
[10]K. Van Den Abeele, P. Johnson, and A. M. Sutin, "Non-linear elastic wave spectroscopy (news) techniques to discern material damage. Part i: Nonlinear wave modulation spectroscopy," Res. Nondestruct. Eval. **12**(1), 17–30 (2000).
[11]K. Van Den Abeele, J. Carmeliet, A. Sutin, and P. Johnson, "Micro-damage diagnostics using nonlinear elastic wave spectroscopy (news)," NDT & E Int. **34**, 239–248 (2001).
[12]G. Montaldo, D. Palacio, M. Tanter, and M. Fink, "Time reversal kaleidoscope: A smart transducer for three-dimensional ultrasonic imaging," Appl. Phys. Lett. **84**, 3879–3881 (2004).
[13]A. M. Sutin, P. A. Johnson, and J. A. TenCate, "Development of nonlinear time reverse acoustics (nltra) for applications to crack detection in solids," in *Proceedings of the 5th World Congress on Ultrasonics* (2003), pp. 121–124.
[14]A. M. Sutin and P. A. Johnson, "Nonlinear elastic wave nde ii. nonlinear wave modulation spectroscopy and nonlinear time reversed acoustics," Rev. Prog. Quant. Nondestr. Eval., edited by Thompson, D. and Chimenti D., **248**, 377–384 (2005).
[15]C. Draeger and M. Fink, "One-channel time reversal of elastic waves in a chaotic 2d-silicon cavity," Phys. Rev. Lett. **79**, 407–410 (1997).
[16]C. Draeger, J.-C. Aime, and M. Fink, "One-channel time-reversal in chaotic cavities: Experimental results," J. Acoust. Soc. Am. **105**, 618–625 (1999).
[17]A. M. Sutin, J. A. TenCate, and P. A. Johnson, "Single-channel time reversal in elastic solids," J. Acoust. Soc. Am. **116**, 2779–2784 (2004).
[18]T. J. Ulrich and R. A. Guyer, "Dispacement field reciprocity in linear and nonlinear reverberant cavities: an experimental exploration with implications for time reversed acoustics," in review (2006).
[19]P. A. Johnson and A. M. Sutin, "Slow dynamics and anomalous nonlinear fast dynamics in diverse solids," J. Acoust. Soc. Am. **117**, 124–130 (2005).

On analysis of exponentially decaying pulse signals using stochastic volatility model

C. M. Chan
Hong Kong Community College, The Hong Kong Polytechnic University, Hong Kong, China

S. K. Tang
Department of Building Services Engineering, The Hong Kong Polytechnic University, Hong Kong, China

H. Wong
Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

(Received 9 August 2005; revised 3 January 2006; accepted 4 January 2006)

A stochastic volatility model incorporating the exponential power distributions is adopted in the present study to analyze exponentially decaying pulses in the presence of background noises of various magnitudes. The discussions are focused on its effectiveness in the determination of the instant of the pulse initiation and the decay constant. The results are compared with those obtained by the conventional short-time Fourier transform. It is found that the present stochastic volatility model can retrieve the instant of the pulse initiation and the decay constant within engineering tolerance even when the noise is slightly stronger that the pulse amplitude. Its performance is substantially better than that of the Fourier transform when the frequency of the decay pulse fluctuates. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2168415]

PACS number(s): 43.60.Uv, 43.60.Jn, 43.60.Cg, 43.60.Hj [EJS] Pages: 1519–1526

I. INTRODUCTION

Quick and accurate signal detection is one of the important research areas in acoustics and building services engineering. Modern buildings are usually heavily serviced and the signals from the services equipments, both the acoustic and vibration signals, are commonly used in the equipment health monitoring processes.[1] Since most of these signals are not stationary, signal processing techniques, such as the short-time Fourier transform (STFT)[2] and the wavelet transforms (WTs),[3] have been proposed and tested in the past few decades. The more efficient extraction of localized signal information using the WT has attracted the interests of many engineers and scientists, especially in the detection of gear faults (for instance, Ref. 4). Pulses and their decays have also been used to measure the energy dissipation of a system. Typical examples are the estimation of room sound absorption and structural damping using reverberation time measurements.[5,6] Under a sufficiently diffused condition, this kind of pulse decay is theoretically an exponential function of time.

In reality, there are background noises, and thus the signal-to-noise ratios (S/N) are of prime importance in the resolution of changes in a signal. For the reverberation time measurement in a noisy environment, the maximum length sequence[7] gives satisfactory results provided that the signal is not too weak when compared to the noise. At high S/N, both the STFT and the WT perform well in the measurement of the decay constant.[8] Wong *et al.*[9] has investigated the use of the WT in locating jumps in a time series when the magnitudes of the jumps are significant.

Conventional parametric models, such as the autoregressive moving averages,[10] have also been used to analyze nonstationary time series. However, these methods do not allow the conditional variances of the data to be a deterministic function of the past observations. This is not realistic in practice, especially when the responses of an excited system are concerned.

A sharp pulse followed by a decay embedded in a random noise, provided that its magnitude is not too weak when compared to that of the noise, will create a relatively higher volatility in the signal fluctuations at the instant the pulse is introduced. A more advanced parametric model, namely the stochastic volatility (SV) models,[11] which have been widely used in modeling time series volatilities, are expected to be useful in the modeling of the above-mentioned pulse signal. The properties of these models and their formulation in the present study will be discussed in the next section. Recently, these models have been used together with efficient Bayesian computational technique in analyzing time series in economic studies (for instance, Ref. 12). However, they are rarely applied, at least to the knowledge of the authors, to deal with engineering problems, though some sort of Bayesian model selection has been considered in the study of geoacoustic signal.[13]

In this paper, the performance of a specific SV model on retrieving the properties of exponentially decaying pulses in the presence of random noises will be presented. It is hoped that the results will provide useful information for the future enhancement of signal detection and machine diagnosis.

II. STOCHASTIC VOLATILITY MODEL

The SV model formulates the volatility by a numerical process which allows the latter to vary stochastically.[11] It involves the use of a statistical distribution so chosen to fit the time series/signal to be analyzed. Both the Guassian distribution and the Student t distribution have been employed

in determining outlying observations in financial time series.[14,15] Recently, Choy and his co-workers have investigated the properties of the SV model which incorporates the exponential power (EP) distributions.[14–16] In their studies, the EP distribution is expressed as scale mixtures representation, which can highly decrease the computational time in the numerical simulation study. Also, the mixing parameters in the representation can be used to identify the extreme data very accurately. Therefore, the EP distributions are adopted in the present study. A brief introduction of the SV model is given in the Appendix.

A. Exponential power distributions

The exponential power family of distributions provides both heavier- and/or lighter-tailed distributions than the normal Guassian one. Let $\{y\}$ be the data set, θ its mean, σ its scale parameter, and $\beta \in (0,2]$ the kurtosis parameter that controls the thickness of the tails. The density function of the EP distribution $\mathrm{EP}(y,\theta,\sigma,\beta)$ is given by[17]

$$\mathrm{EP}(y|\theta,\sigma,\beta) = \frac{1}{2^{(1+\beta/2)}\Gamma(1+\beta/2)\sigma} \exp\left(-\frac{1}{2}\left|\frac{y-\theta}{\sigma}\right|^{2/\beta}\right), \tag{1}$$

where Γ is the gamma function. The corresponding mean and variance are equal to θ and $2^{\beta}\Gamma(3\beta/2)/[\sigma^2\Gamma(\beta/2)]$, respectively. The EP distribution has been studied thoroughly by a number of researchers for statistical modeling and Bayesian robustness (for instance, Choy and Walker[18] and Choy and Chan[15]). Choy and Smith[16] adopted the normal scale mixtures property of the EP density for Bayesian inference using Markov chain Monte Carlo methods with $1<\beta \leqslant 2$. Recently, Walker and Gutierrez-Pena[19] discovered the following uniform scale mixtures representation for the EP density:

$$\mathrm{EP}(y|\theta,\sigma,\beta) = \int_0^{\infty} U(y|\theta-\sigma u^{0.5\beta}, \theta+\sigma u^{0.5\beta}) \times G(u|1+0.5\beta,0.5)\, du, \tag{2}$$

where $U(y|a,b)$ is the uniform density function defined on the interval (a,b) and $G(y|c,d)$ is the gamma density function with mean c/d. This representation is valid for the entire range of β and also allows rewriting the EP distribution into the following hierarchical form:

$$y|u \sim U(\theta-\sigma u^{0.5\beta}, \theta+\sigma u^{0.5\beta}) \quad \text{and} \quad u \sim G(1+0.5\beta,0.5), \tag{3}$$

where u is referred to as the mixing parameter of the uniform scale mixtures representation. It should be noted that the normal and the Laplace distributions are special cases of the EP family with $\beta=1$ and 2, respectively.

B. Bayesian EP SV model

In this paper, the family of EP distributions is considered as a generalization of the normal family to model the signal data. This family provides both leoptkutic and platykurtic shapes of distributions which the normal, Student t, and stable families cannot offer. From a practical point of view, the EP distribution is believed to be appropriate to model certain types of data and it is worthwhile to develop efficient methods for statistical analysis. A Gibbs sampling approach using the uniform scale mixtures is discussed in Sec. III.

The usual choice of the normal distribution for the white noise ε_t of the SV model is replaced by the EP distribution with known kurtosis parameter β:

$$y_t|h_t \sim \mathrm{EP}(0, e^{0.5h_t}, \beta), \tag{4}$$

where t is the index and $t=1,2,\ldots,n$. h is the log volatility. Expression (4) can be rewritten into the following hierarchical form:

$$y_t|h_t,u_t \sim U(-e^{0.5h_t}u_t^{0.5\beta}, e^{0.5h_t}u_t^{0.5\beta}) \quad \text{and} \quad u_t \sim G(1+0.5\beta,0.5), \tag{5}$$

where u is the mixing parameter of the SV model. Normality assumption is still valid for the conditional and marginal distributions of h in the present study:

$$h_t|h_{t-1},\phi,\sigma^2 \sim N(\phi h_{t-1},\sigma^2) \quad \text{and} \quad h_t|\phi,\sigma^2 \sim N(0,\sigma^2/(1-\phi^2)), \tag{6}$$

where $N(0,\alpha)$ is the normal distribution with mean 0 and standard deviation α, and ϕ is the persistence parameter. The above SV model with EP white noise and normal log volatility is referred as the EP-N SV model. In order to complete a full Bayesian framework for the SV model, the following are assigned prior to other model parameters:

$$\sigma^2 \sim G_{inv}(a_\sigma,b_\sigma) \quad \text{and} \quad \phi+1 \sim 2Be(a_\phi,b_\phi), \tag{7}$$

where $Be(a,b)$ is the beta distribution with mean $a/(a+b)$, G_{inv} is the inverse gamma distribution, and a_σ, b_σ, a_ϕ, and b_ϕ are prespecified constants. The prior distribution for ϕ is a shifted beta distribution with density

$$p(\phi) \propto (1+\phi)^{a_\phi-1}(1-\phi)^{b_\phi-1}, \quad |\phi|<1. \tag{8}$$

III. GIBBS SAMPLER FOR EP-N SV MODELS

The simulation-based Gibbs sampling approach[20] is one of the standard methods for carrying out statistical analysis of complicated Bayesian models. The Gibbs sampler allows us to study posterior characteristics via a sequence of iteratively simulated values drawn from a system of full conditional distributions. The efficiency of the Gibbs sampler can be substantially increased if the required samples are drawn from distributions of some standard forms. Gibbs sampling has also been used in underwater acoustic application.[21]

The joint distribution of $\vec{y}=(y_1,y_2,\ldots,y_n)$, $\vec{h}=(h_1,h_2,\ldots,h_n)$, $\vec{u}=(u_1,u_2,\ldots,u_n)$, ϕ, and σ^2 is given by

$$p(\vec{y},\vec{h},\vec{u},\phi,\sigma^2) = \prod_{t=1}^{n} p(y_t|h_t,u_t)p(h_1|\phi,\sigma^2)p(\vec{u})p(\phi)p(\sigma^2). \tag{9}$$

Then, the Gibbs sampling scheme performs successive random variate generation from the following conditional distri-

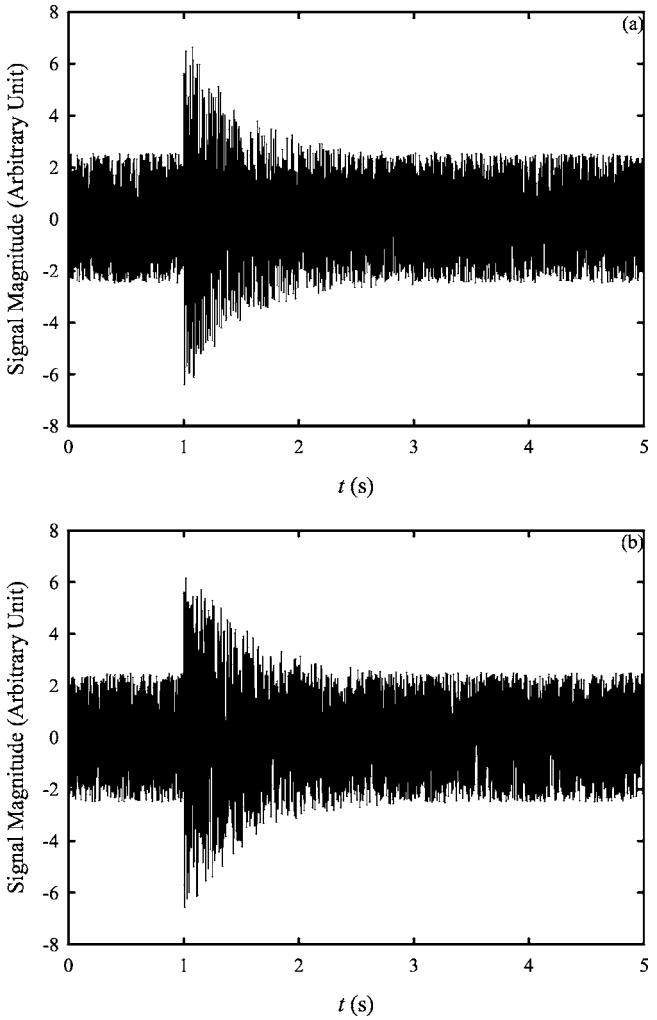


FIG. 1. Examples of exponentially decaying signals with $S/N=10$ dB. (a) Constant f. (b) Time varying f.

butions. The duration of the Gibbs sampler computation varies from a few minutes to about 25 min on a Pentium V personal computer, depending on the kurtosis parameter of the EP distribution. The number of iterations adopted is 12 000 with the first 2000 iterations as the burn-in period. In the foregoing discussions, $\vec{h}_{-t}=(h_1,\ldots,h_{t-1},h_{t+1},\ldots,h_n)$ and $\vec{u}_{-t}=(u_1,\ldots,u_{t-1},u_{t+1},\ldots,u_n)$.

A. Full conditional densities of *h*

Full conditional density of h is given by

$$p(h_t|\vec{y},\vec{h}_{-t},\vec{u},\phi,\sigma^2) \propto p(y_t|h_t,u_t)p(h_t|h_{t-1},u_t,\phi,\sigma^2)p(h_{t+1}|h_t,u_{t+1},\phi,\sigma^2). \tag{10}$$

One can then show that these full conditional distributions are truncated normal of the form

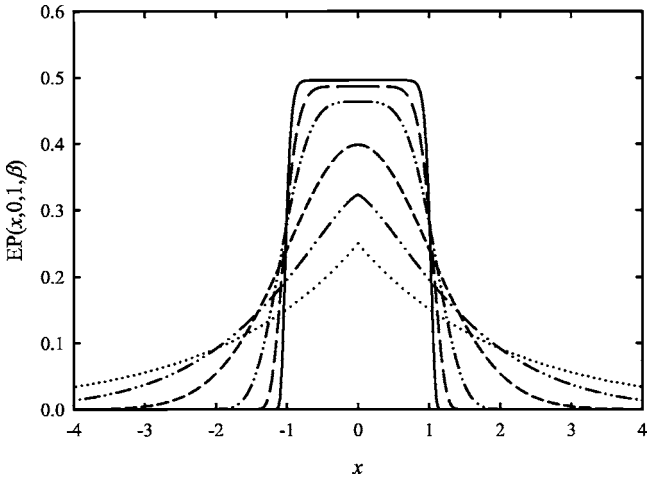


FIG. 2. Effects of β on the shape of an EP distribution. ———: $\beta=0.1$; – – –: $\beta=0.25$; –·–: $\beta=0.5$;- - -:$\beta=1$; –·–·–: $\beta=1.5$; ···: $\beta=2$.

$$h_t|\vec{y},\vec{h}_{-t},\vec{u},\phi,\sigma^2 \sim \begin{cases} N(\phi h_{t+1}-\sigma^2/2,\sigma^2), & t=1, \\ N\left(\dfrac{\phi(h_{t-1}+h_{t+1})-\sigma^2/2}{1+\phi^2},\dfrac{\sigma^2}{1+\phi^2}\right), & 2\leqslant t\leqslant n-1, \\ N(\phi h_{t-1}-\sigma^2/2,\sigma^2), & t=n, \end{cases} \tag{11}$$

subject to $h_t>\ln y_t^2-\ln\beta^2-\alpha\ln u_t$. The algorithm proposed by Robert[22] is an efficient method for generating random variates from the truncated normal distribution.

B. Full conditional densities of *u* and σ^2

By representing the EP density into a uniform scale mixtures form, one can show that the full conditional distribution of the mixing parameter u is a truncated exponential distribution of the form

$$u_t|\vec{y},\vec{h},\vec{u}_{-t},\phi,\sigma^2 \sim \exp(u|0.5)=0.5e^{-0.5u}, \tag{12}$$

subject to the condition

$$u_t^{\beta}>\frac{y_t^2}{\beta^2}e^{-h_t}. \tag{13}$$

The inversion method can be used to sample random variates from the truncated exponential distribution. For σ^2, the use of conjugate prior leads to an inverse gamma full conditional distribution and σ^2 can then be directly sampled from

$$\sigma^2|\vec{y},\vec{h},\vec{u},\phi \sim G_{inv}\left(a_\sigma+\frac{n}{2},b_\sigma+\frac{1}{2}\left((1-\phi^2)h_1^2+\sum_{t=2}^{n}(h_t-\phi h_{t-1})^2\right)\right). \tag{14}$$

C. Full conditional densities of ϕ

The full conditional density of ϕ is given by

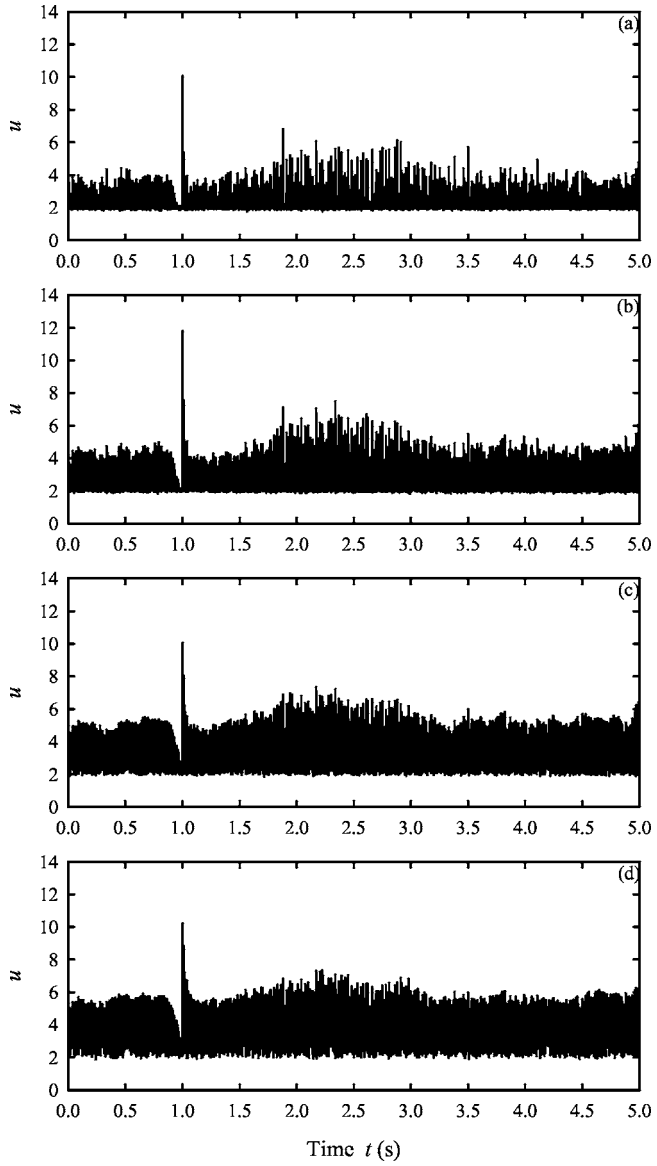


FIG. 3. Effects of β on the time variation of the mixing parameter u. (a) $\beta=0.25$; (b) $\beta=0.75$; (c) $\beta=1.5$; (d) $\beta=2.0$. S/N=10 dB.

$$p(\phi|\vec{y},\vec{h},\vec{u},\sigma^2) \propto p(h_1|\phi,\sigma^2)\prod_{t=1}^{n} p(h_t|h_{t-1},\phi,\sigma^2)p(\phi). \tag{15}$$

It can easily be verified that $\Pi_{t=2}^{n} p(h_t|h_{t-1},\phi,\sigma^2)$ is proportional to a normal density of ϕ with mean μ_ϕ and variance σ_ϕ^2 where

$$\mu_\phi = \frac{\sum_{t=1}^{n-1} h_t h_{t+1}}{\sum_{t=1}^{n-1} h_t^2} \quad \text{and} \quad \sigma_\phi^2 = \frac{\sigma^2}{\sum_{t=1}^{n-1} h_t^2}. \tag{16}$$

One can express the full conditional density by a product of a truncated normal function and a shifted beta density function:

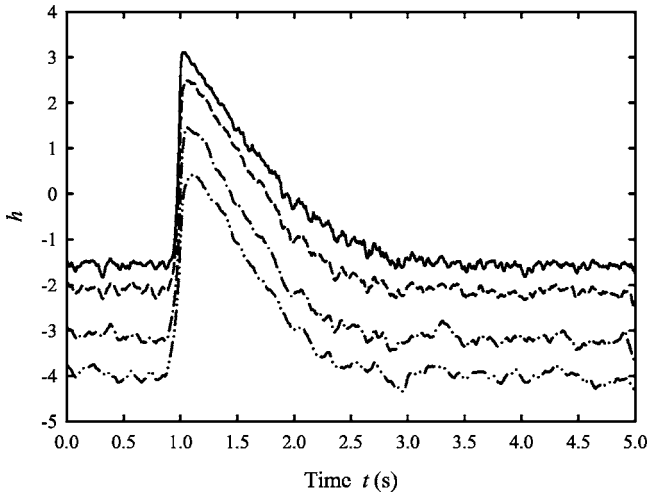


FIG. 4. Effects of β on the time variation of the log-volatility h. ———: $\beta=0.25$; - - -:$\beta=0.75$; –·–: $\beta=1.5$; –·· –: $\beta=2.0$. S/N=10 dB.

$$p(\phi|\vec{y},\vec{h},\vec{u},\sigma^2) \propto N\left(\frac{\sum_{t=1}^{n-1} h_t h_{t+1}}{\sum_{t=2}^{n-1} h_t^2}, \frac{\sigma^2}{\sum_{t=2}^{n-1} h_t^2}\right) \times (1+\phi)^{a_\phi - 1/2}(1-\phi)^{b_\phi - 1/2} \tag{17}$$

for $|\phi<1|$. Sampling random variates from this full conditional density can be done easily by using the Metropolis-Hastings method.[23]

IV. NUMERICAL EXAMPLES

An exponentially decaying harmonic wave is chosen for the illustrations. This kind of signal can be regarded as the simplest (but important) signal in acoustics and vibration studies. The noise signals $n(\tau)$, where τ denotes measurement time, used in the foregoing illustrations are white noises with zero mean values and are Gaussian distributed fluctuations. The combined signal $y(\tau)$ is

$$y(\tau) = s(\tau) + n(\tau), \tag{18}$$

where $s(\tau)$ contains the pulses. The S/N in dB is defined as

$$\text{S/N} = 10\log_{10}(|s|_{\max}/|n|_{\max}). \tag{19}$$

Suppose the pulse is initiated at $\tau=\tau_o$ and let η be the decay constant; the signal $y(\tau)$ is

$$y(\tau) = e^{-\eta(\tau-\tau_o)}\cos(2\pi f(\tau-\tau_o))\mathrm{H}(\tau-\tau_o) + n(\tau), \tag{20}$$

where H denotes the Heavside step function and f is the frequency of the harmonic wave. The present investigation focuses on the determination of the instant of the pulse initiation and the decay constant η. The S/N ranges from $+\infty$ to -10 dB. Without loss of generality, η is fixed at 2 and f is normally set at 50 Hz, but is allowed to have a narrow ±10% time fluctuation in some cases discussed later. τ_o =1 s throughout the present investigation. Figures 1(a) and 1(b) illustrate the hypothetical signals with S/N =3 dB with constant $f(\equiv 50$ Hz) and time varying f(=50 Hz+10% time fluctuation), respectively. One can notice that the signal with fluctuating f does not show a clearly exponential decay feature [Fig. 1(b)].

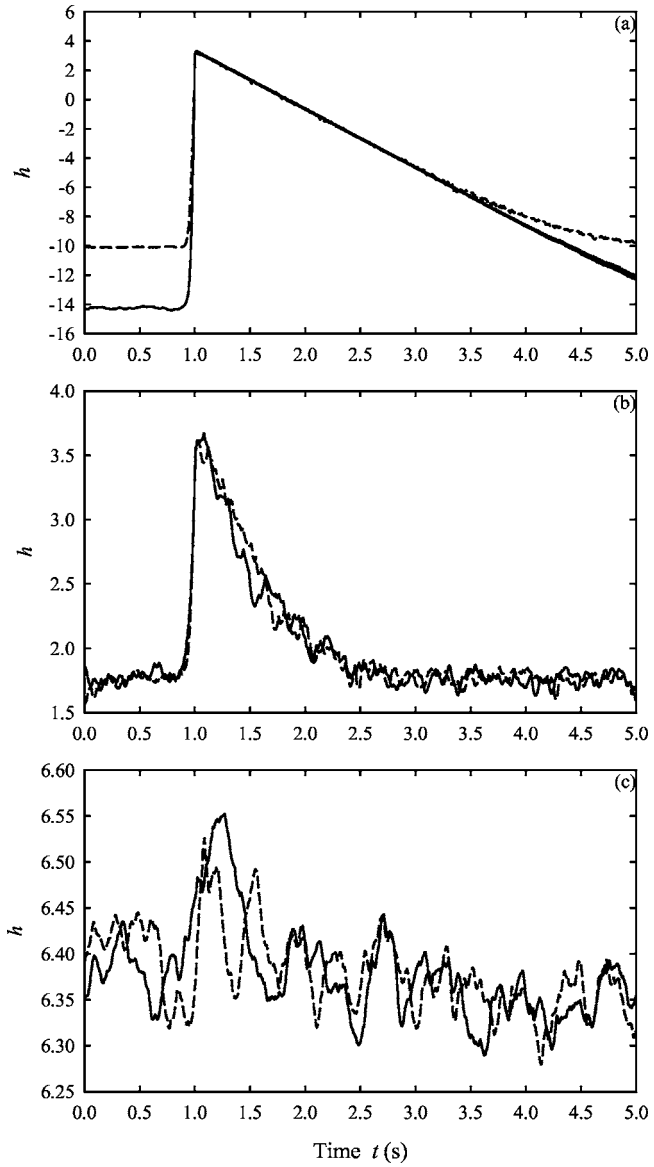


FIG. 5. Time variations of h at increased noise magnitude with $\beta=0.1$. (a) S/N=$+\infty$ dB; (b) S/N=3 dB; (c) S/N=−7 dB———: constant f; – –: time varying f.

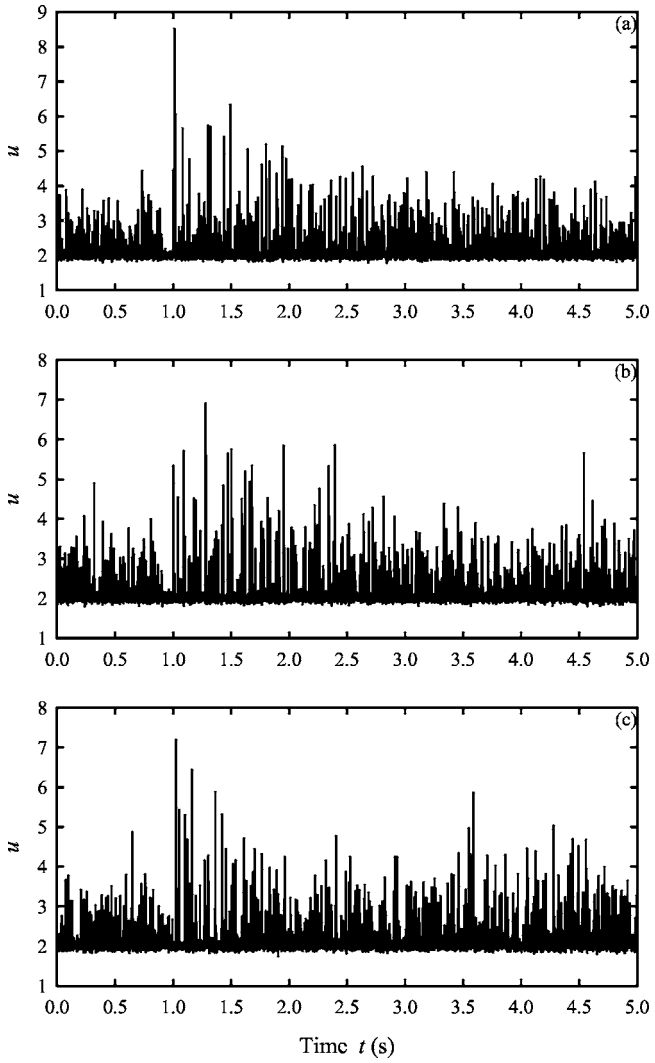


FIG. 6. Time variations of u at increased noise magnitude with $\beta=0.1$. (a) S/N=3 dB; (b) S/N=0 dB; (c) S/N=−7 dB.

The parameter β in the SV model determines the shapes of the EP distributions and thus has significant impact on the modeling of the decaying signal by the SV model. The EP distributions at various β are given in Fig. 2. A large β gives rise to a thick tail of the distribution and is in general not good for detecting changes as a thick tail distribution tends to down-weight the extreme values and thus affect the detection of pulses in the presence of a random noise.

Two parameters are important in the SV model. They are the mixing parameter u and the log volatility h. The former illustrates fluctuation and thus should be able to suggest the instants of rapid changes in a signal. The latter should follow the shapes of the decaying pulses. In the analysis of a decaying pulse, these two parameters should not be considered in isolation.

A. Constant *f*

Figure 3 shows the time variations of u at different β for S/N=10 dB. The signal is considerably stronger than the background noise in this case. One can notice a prominent sharp peak within a continuous background spikes in each case illustrated. This peak appears around the instant of the pulse initiation. In fact, the magnitudes of the background spikes before the peak, which are due to the background Gaussian noise, relative to that of the sharp peak increase with increasing β. It is found from a closer look at the data at around $\tau=1$ s that the magnitudes of the spikes continue to decrease when τ increases towards 1 s as shown in Fig. 3(a) with S/N=10 dB and $\beta=0.25$. A clear reversal is observed

TABLE I. Estimated τ_o and η (constant f).

S/N (dB)	SV model		STFT
	τ_o (S)	η (s^{-1})	η (s^{-1})
$+\infty$	1.000	2.00	2.00
10	1.000	2.05	2.05
3	1.000	2.16	1.81
0	1.001	2.42	1.58
−3	1.002	2.36	1.30
−7	1.022	⋯	⋯

TABLE II. Estimated τ_o and η (fluctuating f).

	SV model		STFT	
S/N (dB)	τ_o (S)	η (s^{-1})	t_o (s)	η (s^{-1})
∞	1.000	2.00	0.96	2.53
10	1.000	2.04	0.96	2.83
3	1.000	2.30	⋯	1.85
0	1.010	2.48	⋯	1.10
−3	1.001	2.37	⋯	⋯
−7	⋯	⋯	⋯	⋯

at $\tau=1$ s. A similar phenomenon is also observed at other values of β investigated in the present study under this signal-to-noise ratio.

The time variations of the log volatility h for S/N $=10$ dB with various β are presented in Fig. 4. The result obtained with $\beta=0.1$ resembles very much that with $\beta=0.25$ and thus is not presented. The patterns of the h variations in general follow a linear decay but the increase in β appears to have smoothed the pulse identity. The linear decay also becomes less obvious as β increases. Therefore, only results at $\beta=0.1$ will be presented in the foregoing discussions.

A decrease in the S/N results in a rocky decay of h and the exponential decay becomes slightly untraceable at S/N $=-7$ dB though a decaying pulse is still suggested by the time variation of h (Fig. 5). However, a clear reversal in u at $\tau\sim 1$ s remains prominent even up to a S/N of 0 dB and some indications of an abrupt signal jump are still observable at S/N$=-7$ dB (Fig. 6). One should note that the instant of the peak u does not necessarily collapse with that of the pulse initiation. The focus should be on the instant of the reversal. It should also be noted that the magnitude of the u peak does not carry much meaning in engineering applications. Table I summarizes the instant of the u reversal under different S/Ns with $\beta=0.1$. Though one can anticipate ambiguity in determining the instant of this major u reversal at strong background noise magnitude and thus the error in locating the instant of the pulse initiation, the prominent sharp rise of u, such as that observed in Fig. 6(c), indicates together with the variation of h that some important changes are embedded in a random signal. This shows the versatility of the present SV model in detecting changes.

Apart from detecting the instant of the pulse initiation, the determination of the decay constant is also an important task in the signal analysis.[8] The method shown in the Appendix is adopted to minimize the effects of the background noise in the process. The average h for the background noise can be obtained using 800 data points starting from $\tau=0$ s. Data points close to the major u reversal should be avoided. The patterns of h variations after the background noise correction are very similar to those of the impulse decays inside rooms and thus are not presented. The decay constant η's estimated under different S/Ns are shown in Table I. η's from the SV model are obtained using the 1000 data points after the major u reversal and the correlation coefficients R^2 of the regression are in general greater than 0.9 for S/N$\geqslant -3$ dB. The corresponding values obtained from the STFT with a frequency resolution of 3.9 Hz and 60% data overlapping are also presented in Table I for the sake of comparison. Allowing for error in the estimation of the slopes of the decay curves, the two sets of results are comparable. However, the STFT is not able to provide the time resolution one can obtain from the present SV model unless the frequency resolution is lowered to 40 Hz and the S/N is high. Thus, the STFT is not suitable for the determination of τ_o.

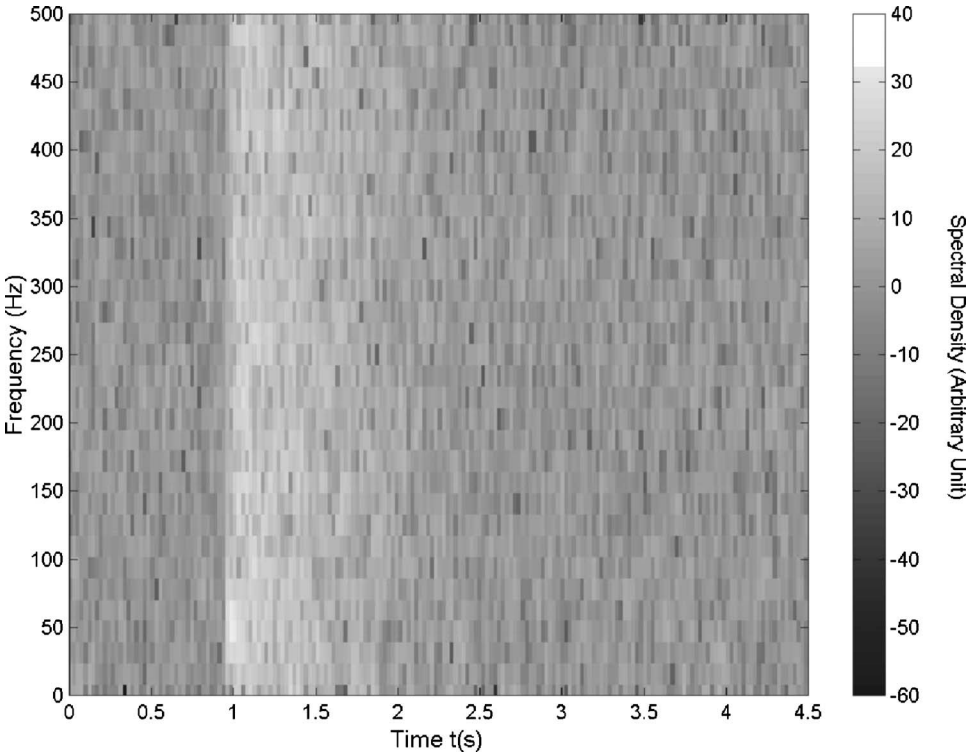


FIG. 7. Time-frequency plot of an exponential decaying signal with time varying f. S/N$=10$ dB, frequency resolution: 15.6 Hz.

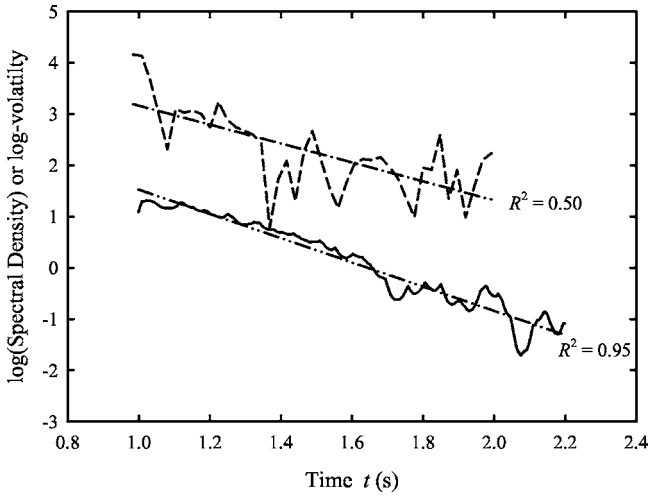


FIG. 8. Decay curves with time varying f at S/N=3 dB———: log-volatility $h(\beta=0.1)$–·–·–: regression line for h; – –:log (spectral density); –·–: regression line for log (spectral density).

B. Time varying *f*

The introduction of a ±10% Gaussian fluctuation in f does not greatly affect the pattern of the time varying mixing parameter u. A major reversal of u can again be observed at the instant close or equal to τ_o when S/N>-3 dB, but such reversal becomes ambiguous when S/N is further decreased (not presented here). The fluctuation in f does increase the log volatility h when the background noise is strong relative to the signal as shown in Fig. 5. The τ_o and η determined from the present SV model with EP distribution having $\beta=0.1$ are tabulated in Table II. The randomness in f does not produces observable deterioration of the SV model performance, except that the error in locating τ_o becomes slightly higher than that in the constant f case when S/N drops below 0 dB.

The randomness in f, however, results in a broadband time-frequency distribution as computed by the STFT (Fig. 7). In order to cover the range of the frequency fluctuation, the frequency resolution here is taken to be 15.6 Hz. Sixty percent data overlapping is again adopted in the STFT and the time resolution is 0.024 s. One can notice from Table II that the STFT does not provide reliable estimation of η. The ambiguity of τ_o determination is high when S/N is less than 3 dB. Though η is estimated to be 1.85 by STFT at S/N=3 dB, the actual decay curve is problematic as shown in Fig. 8 (all corrected for background noise). The linear decay is not so reflected by the STFT, while that obtained from the present SV model is still acceptable. Reasonably good linear regression between h and τ is also obtained at S/N=0 and −3 dB (not shown here), but the STFT is unable to reveal a decay trend when S/N decreases below 0 dB.

V. CONCLUSIONS

A stochastic volatility model incorporating the exponential power distribution is used in the present study to capture the initiation and decay of exponentially decaying signals in the presence of random background noises. Its performance is compared with that of the short-time Fourier analysis.

When the pulse is a decaying harmonic wave of constant frequency, the accuracy of decay constant estimation using the present stochastic volatility model is very good when the signal-to-noise ratio is not less than 3 dB. Such accuracy deteriorates as the signal-to-noise ratio decreases, but it is still comparable to that of the short-time-Fourier transform. The present model is able to detect the instant of the pulse initiation even up to a signal-to-noise ratio of −7 dB within engineering tolerance. The short-time Fourier transform does not provide a comparable time resolution for such detection unless the frequency resolution is scarified and the signal-to-noise ratio is high.

The introduction of a time varying frequency (±10%) does not produce significant effect on the performance of the stochastic volatility model, though it raises up the volatility when the background noise is relatively strong, making the detection of the pulse difficult once the signal-to-noise ratio drops below 0 dB. However, the performance of the short-time Fourier transform is substantially worsened even in the absence of the background noise.

ACKNOWLEDGMENT

CMC is supported by a staff development porgramme of the Hong Kong Community College, The Hong Kong Polytechnic University.

APPENDIX: SV MODELING AND DECAY CONSTANT ESTIMATION

In the present study, a hypothetical signal y made up of an exponential decaying signal d initiated at $\tau=\tau_o(>0)$ and a continuous white noise (normally distributed) n is considered. The SV model suggests[11]

$$y = d + n = H_t^{1/2}\varepsilon_t, \tag{A1}$$

where ε_t is a time fluctuating white noise and H_t the volatility. Suppose one can find $d=H_{d,t}^{1/2}\varepsilon_t$ (no noise) and $n=H_{n,t}^{1/2}\varepsilon_t$, one then have the following approximate relationship.

$$H_t^{1/2} = H_{d,t}^{1/2} + H_{n,t}^{1/2}. \tag{A2}$$

Let $h_t=\log H_t$, $h_{d,t}=\log H_{d,t}$ and $h_{n,t}=\log H_{n,t}$, one can then find

$$h_{d,t} = 2\log(e^{h_t/2} - e^{h_{n,t}/2}). \tag{A3}$$

At $\tau>\tau_o$ and excluding the sinusoidal time fluctuation, one obtains $|d| \sim Ae^{-\alpha(\tau-\tau_o)}$ and $H_{d,t}^{1/2} \sim Ae^{\alpha\tau_o}e^{\alpha\tau}$ so that $h_{d,t}=\log H_{d,t}=-2\alpha\tau+2\log(Ae^{\alpha\tau_o})$, and

$$\log(Ae^{\alpha\tau_o}) - \alpha\tau = \log(e^{h_t/2} - e^{h_{n,t}/2}). \tag{A4}$$

Since n is normally distributed and thus can be treated as ε_t multiplied by a constant, it is straight-forward to conclude that

$$n = A(\mathrm{S/N})\varepsilon_t \Rightarrow e^{h_{n,t}/2} \sim A \times \mathrm{S/N}. \tag{A5}$$

One can normalize y by A such that $A=1$ in the above derivation. $h_{n,t}$ can be obtained from the data at $\tau<\tau_o$. Usually, an average from less than 400 points is good enough to get a reliable estimate of $h_{n,t}$ when S/N$\geqslant 0$.

[1]R. B. Randall, *Frequency Analysis* (Brüel & Kjær, Nærum, Denmark, 1987).
[2]C. H. Hodges, J. Power, and J. Woodhouse, "The use of sonogram in structural acoustics and an application to the vibrations of cylindrical shells," J. Sound Vib. **101**, 203–218 (1985).
[3]I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," IEEE Trans. Inf. Theory **36**, 961–1005 (1990).
[4]P. D. McFadden and W. J. Wang, "Early detection of gear failure by vibration analysis—I. Calculation of the time-frequency distribution," Mech. Syst. Signal Process. **7**, 193–203 (1993).
[5]H. Kuttruff, *Room Acoustics* (E & FN Spon, New York, 2000).
[6]M. Heckl, "Measurement of absorption coefficients on plates," J. Acoust. Soc. Am. **34**, 803–808 (1962).
[7]M. Vorländer and M. Kob, "Practical aspects of MLS measurements in building acoustics," Appl. Acoust. **52**, 239–258 (1997).
[8]S. K. Tang, "On the time-frequency analysis of signals that decay exponentially with time," J. Sound Vib. **234**, 241–258 (2000).
[9]H. Wong, W. C. Ip, and Y. Li, "Detection of jumps by wavelets in a heteroscedastic autogressive model," Stat. Probab. Lett. **52**, 365–372 (2001).
[10]G. E. P. Box and G. M. Jenkins, *Time Series Analysis Forecasting and Control*, 2nd ed. (Holden-Day, San Francisco, 1976).
[11]S. J. Taylor, "Modelling stochastic volatility," Math. Finance **4**, 183–204 (1994).
[12]E. Jacquier, N. G. Polson, and P. E. Rossi, "Bayesian analysis of stochastic volatility models (with discussion)," J. Bus. Econ. Stat. **12**, 371–417 (1994).
[13]D. J. Battle, P. Gerstoft, W. S. Hodgkiss, and W. A. Kuperman, "Bayesian model selection applied to self-noise geoacoustic inversion," J. Acoust. Soc. Am. **116**, 2043–2056 (2005).
[14]S. T. B. Choy and C. M. Chan, "Scale mixtures distributions in insurance applications," Austin Bull. **33**, 93–104 (**2003**).
[15]S. T. B. Choy and C. M. Chan, "Bayesian student-*t* stochastic volatility models via two-stage mixtures representation," 20th International Workshop on Statistical Modelling. Sydney, Australia, 10–15 July 2005.
[16]S. T. B. Choy and A. F. M. Smith, "Hierarchical models with scale mixtures of normal distribution," TEST **6**, 205–211 (1997).
[17]G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis* (Wiley, New York, 1992).
[18]S. T. B. Choy and S. G. Walker, "The extended exponential power distribution and Bayesian robustness," Stat. Probab. Lett. **65**, 227–232 (2003).
[19]S. G. Walker and E. Gutierrez-Pena, in *Bayesian Statistics* (Oxford U. P., Oxford, 1999), Vol. **6**, pp. 685–710.
[20]A. E. Gelfand, S. E. Hills, A. Racine-Poon, and A. F. M. Smith, "Illustration of Bayesian inference in normal modes using Gibbs sampling," J. Am. Stat. Assoc. **85**, 972–985 (1990).
[21]Z. H. Michalopoulou and M. Picarelli, "Gibbs sampling for time-delay and amplitude estimation in underwater acoustics," J. Acoust. Soc. Am. **117**, 799–808 (2005).
[22]C. P. Robert, "Simulation of truncated normal variables," Stat. Comput. **5**, 121–125 (1995).
[23]S. Kim, N. Shephard, and S. Chib, "Stochastic volatility: Likelihood inference and comparison with ARCH models," Metalloberflache **65**, 361–393 (1998).

The relation between electrophysiologic channel interaction and electrode pitch ranking in cochlear implant recipients[a)]

Michelle L. Hughes[b)] and Paul J. Abbas
Department of Otolaryngology–Head and Neck Surgery, University of Iowa, 200 Hawkins Dr., Iowa City, Iowa 52242 and Department of Speech Pathology and Audiology, University of Iowa, Iowa City, Iowa 52242

(Received 31 May 2005; revised 9 December 2005; accepted 12 December 2005)

The primary goal of this study was to examine the relation between electrophysiologic measures of channel interaction and the ability to discriminate pitch between electrodes in a psychophysical pitch-ranking task. It was hypothesized that cochlear implant recipients should perform better on an electrode pitch-ranking task when using electrodes with less channel interaction as measured with the electrically evoked compound action potential (ECAP). The width of the ECAP channel interaction function was compared with the slope of the pitch-ranking function for 10 adult Nucleus 24 recipients. Results showed no significant correlation between electrode pitch-ranking ability and width of the ECAP channel interaction function for individual subjects or for group data. Additionally, there was no significant correlation between speech perception performance and either pitch-ranking ability or width of the ECAP channel interaction function. These results suggest that the width of the ECAP interaction function may not be an accurate predictor of the ability to discriminate between electrodes on the basis of pitch.
 [DOI: 10.1121/1.2163273]

PACS number(s): 43.64.Me, 43.64.Pg, 43.66.Fe, 43.71.Ky [BLM] Pages: 1527–1537

I. INTRODUCTION

Multichannel cochlear implants offer improved frequency selectivity over single channel implants by taking advantage of the tonotopic organization in the cochlea. The result is expanded spectral representation of the speech signal, which has yielded dramatic improvements in speech perception ability (e.g., Gantz *et al.*, 1988). However, to obtain true frequency selectivity as a function of cochlear place, each electrode must stimulate a specific neural population that yields a distinct pitch percept. If two or more electrodes stimulate overlapping neural populations, whether simultaneously or in sequence, then channel interaction may occur. The perceptual result of channel interaction can be changes in loudness (Shannon, 1983, 1985; White *et al.*, 1984) or reduced ability to detect one signal that follows another (e.g., Boex *et al.*, 2003; Chatterjee and Shannon, 1998; Throckmorton and Collins, 1999). Physiologic results of channel interaction include either facilitation or reduction of auditory neural responses for stimuli that are either spatially or temporally close (e.g., Abbas and Brown, 1988; Abbas *et al.*, 2004; Brown *et al.*, 1990; Cartee *et al.*, 2000; Cohen *et al.*, 2003; White *et al.*, 1984). These phenomena have the potential to produce spectral or temporal smearing, which may negatively affect speech perception outcomes with a cochlear implant. The relation between physiologic and psychophysical measures of channel interaction is not well known. Specifically, it is unclear the extent to which physiologic channel interaction within the cochlea affects the ability to discriminate pitch percepts between electrodes.

Channel interaction has been measured physiologically with the electrically evoked compound action potential (ECAP) using a forward masking technique (Abbas *et al.*, 2004; Cohen *et al.*, 2003; Peeters *et al.*, 1998). With this technique, masker and probe stimuli are presented sequentially to different electrodes (typically either masker or probe is fixed on one electrode while the other varies across the array). The resulting ECAP amplitude patterns indicate the amount of spatial overlap between neural populations stimulated by masker and probe electrodes. These ECAP measures therefore characterize the amount of interaction between the physiologic spatial selectivity functions of two electrodes. Results from those studies suggest that masking effects diminish with increased distance between masker and probe electrodes and that the peak of the masking patterns tend to occur with the masker and probe on the same electrode, indicating the greatest amount of interaction.

We might predict that if significant channel interaction were present within the cochlea, then ECAP forward masking patterns would be very broad (i.e., there would be little change in amplitude with greater separation between masker and probe electrodes). Excessive overlap between stimulated neural populations might result in spectral smearing and re-

[a)]Portions of this work were presented in "ECAP channel interaction and electrode discrimination ability in cochlear implants," Proceedings of VIII International Cochlear Implant Conference, Indianapolis, IN, May 2004; "The relationship between electrophysiologic and psychophysical measures of spatial spread in cochlear implants," Proceedings of Scientific and Technology Meeting of the American Auditory Society, Scottsdale, AZ, March 2003; 26th Midwinter Meeting of the Association for Research in Otolaryngology, Daytona Beach, FL, February 2003; and 7th International Cochlear Implant Conference, Manchester, United Kingdom, September 2002.

[b)]Present address: Boys Town National Research Hospital, Lied Learning and Technology Center, 425 North 30th St., Omaha, Nebraska 68131. Electronic mail: hughesml@boystown.org

TABLE I. Demographic data for subjects participating in this study.

Subject number	Device type	Ear	Age at implant	Duration implant use	Duration deafness	Reported etiology of deafness
M15	24M	Right	69 yr	5 yr	10 yr	Infection
M24	24M	Left	51 yr	4.5 yr	48 yr	Trauma
M35b[a]	24M	Left	69 yr	4 yr	1 yr	Noise
M50	24M	Left	72 yr	3 yr	5 yr	Sudden
M54b[a]	24M	Right	52 yr	3 yr	7 yr	Unknown
R9	24R	Right	35 yr	2 yr	2 yr	Unknown
R13	24R	Right	66 yr	2 yr	5 yr	Progressive/hereditary
R15	24R	Right	39 yr	2 yr	3 yr	Hereditary
R21	24R	Right	68 yr	5 mo	10 yr	Unknown
R22	24R	Right	70 yr	5 mo	2 mo	Noise/sudden

[a]Bilaterally implanted.

duced electrode pitch discrimination ability, possibly leading to poorer speech perception ability. Some studies have reported comparable forward masking effects between physiologic and psychophysical measures as a function of temporal (Brown *et al.*, 1996) or spatial (Cohen *et al.*, 2003) separation of masker and probe. These outcomes suggest that physiologic forward masking patterns could potentially be used to predict time-consuming measurement of psychophysical forward masking patterns. Another study found a significant relation between psychophysical forward masking and electrode discrimination ability in adult cochlear implant users (Throckmorton and Collins, 1999). No study has compared physiologic spatial forward masking patterns with electrode pitch discrimination or pitch ranking ability. Last, a number of studies have reported a significant positive relation between electrode discrimination ability and speech perception performance (Busby and Clark, 2000; Henry *et al.*, 2000; Nelson *et al.*, 1995; Throckmorton and Collins, 1999).

The primary goal of this study was to examine the relation between electrophysiologic measures of channel interaction using the ECAP and the ability to discriminate between electrodes on the basis of pitch using a psychophysical pitch-ranking task. It was hypothesized that subjects should perform better on an electrode pitch-ranking task when using electrodes with more selective masking patterns (i.e., less channel interaction) as measured with the ECAP. Similarly, electrodes with broad ECAP masking patterns (more interaction) should be correlated with poorer performance on an electrode pitch-ranking task. If ECAP masking patterns reflect pitch-ranking ability on specific electrodes, then ECAP measures could serve as a quick means to identify spectrally indiscriminable electrodes that could be eliminated from the speech processor program, which may potentially improve speech perception performance.

II. METHODS

A. Subjects

Ten adult cochlear implant recipients participated in this study. Five subjects used the Nucleus 24M straight array and five subjects used the Nucleus 24R(CS) Contour array. Demographic data for participants are listed in Table I. Subjects M35b and M54b were implanted bilaterally; however, the ear indicated in Table I was the only ear tested for that subject in this study. Subject R9 had previously been implanted with an experimental short-electrode cochlear implant (Nucleus CI 6+16+2M) for 9 months prior to being reimplanted with a Contour array in the same ear. Subject M50 had a history of a short circuit between electrodes 12 and 13, although normal impedance was measured for those electrodes at the time of data collection. This subject reported nonauditory percepts when electrodes 12 and 13 were stimulated, so those electrodes were not tested during this study. This subject also had four stiffening rings outside of the cochlea as noted at the time of surgery. Last, electrodes 10 and 11 were not tested for subject R22 due to a short circuit. At the time of testing, all subjects had at least 5 months of experience with their implant.

B. Dynamic range measures

Behavioral dynamic range was measured for each subject to determine stimulus levels for ECAP and electrode pitch ranking measures. Behavioral thresholds (T-levels) and maximum comfort levels (C-levels) were measured on all electrodes in the array using traditional clinical programming procedures. These procedures are described in detail in a companion paper (Hughes and Abbas, 2006). Behavioral levels were measured using the WinDPS programming software (R116, build 445) installed on a desktop computer. The computer controlled a SPrint body-worn speech processor through a Processor Control Interface (PCI). Stimuli consisted of a train of 25 μs/phase biphasic current pulses presented at a rate of 250 pps. Pulse train duration was 500 ms with a 500-ms silent interval between stimulus presentations. Stimulation was presented in monopolar mode relative to the extracochlear ground electrode MP1. Levels were balanced at 50% and 100% of the dynamic range. Subjects were instructed to identify any electrode(s) that seemed out of pitch order when stimuli were swept across electrodes. "Pitch order" was described to patients as sounding like notes going up a piano keyboard. All subjects reported proper ordering of pitch when the sweeps were performed.

C. ECAP measures

1. Stimuli and procedure

ECAP measures were recorded using the commercially available Neural Response Telemetry (NRT, version 3.0) software. Stimuli were delivered through a SPrint body-worn speech processor interfaced with either the Processor Control Interface (PCI) or Portable Programming System (PPS). The following default parameters were used: rate, 80 Hz; pulse width, 25 μs/phase; gain, 60 dB; masker/probe reference electrode, MP1; and recording reference electrode, MP2. Responses were recorded using 50–200 averages depending on overall amplitude and noise level of each waveform.

With NRT, stimulus artifact is removed from the ECAP response using a subtraction method described previously (Abbas *et al.*, 1999, 2004; Brown *et al.*, 1998, 2000; Cohen *et al.*, 2003; Dillier *et al.*, 2002). The subtraction method can also be used to assess interaction between neural populations that are stimulated by different electrodes (Abbas *et al.*, 2004; Cohen *et al.*, 2003). Figure 1 illustrates the effect of masker electrode location on ECAP amplitude. In each of the three panels, trace A shows the ECAP response in the probe-alone condition, and trace B shows the response in the forward-masked (masker-plus-probe) condition, where the first pulse is the masker and the second pulse is the probe. Trace A-B shows a simplified representation of the ECAP when the probe response in trace B is subtracted from the probe response in trace A. For clarity, the two additional traces from the NRT subtraction method (masker-alone and system noise associated with switching of the current source) are not included in Fig. 1. Stimulated electrodes are shown as darkened circles and the gray halo around each stimulated electrode represents hypothetical spread of current from the electrode. Bolded lines indicate stimulated nerve fibers and thin lines indicate nonstimulated fibers. In the top panel of Fig. 1, the masker and probe are applied to the same electrode. In this condition the masker and probe stimulate the same population of nerve fibers (shown in bold). As a result, all fibers discharge in response to the masker, yielding no response to the probe in the forward-masked condition (trace B). In the subtracted trace (A-B), the ECAP amplitude is largest when masker and probe stimulate completely overlapping neural populations. In the middle panel of Fig. 1, the masker is applied to a different electrode than the probe, yielding a partial overlap of stimulated neurons. Stimulated nerve fibers that are common to both the masker and the probe are shown as bold dotted lines. These "common" fibers are recruited first by the masker, leaving a subset of fibers to respond to the probe in the forward-masked condition. The result is a reduced-amplitude ECAP in response to the probe in the forward-masked condition (trace B). In the subtracted trace (A-B), a small ECAP results. In the bottom panel of Fig. 1, the masker and probe are applied to widely spaced electrodes that excite completely separate, nonoverlapping neural populations. Because the stimulated neural populations are independent, the masker has no effect on the fibers recruited by the probe. As a result, an ECAP is measured in response to both masker and probe stimuli (trace B), yielding no response in the subtracted trace (A-B). With this subtraction method, largest ECAP responses are typically obtained when the masker and probe are on the same electrode, and diminished or absent ECAP responses are obtained for larger spatial separations of masker and probe. Lack of an ECAP response in the subtracted waveform indicates no measurable effect of the masker on the probe and a lack of peripheral channel interaction between the masker and probe electrodes.

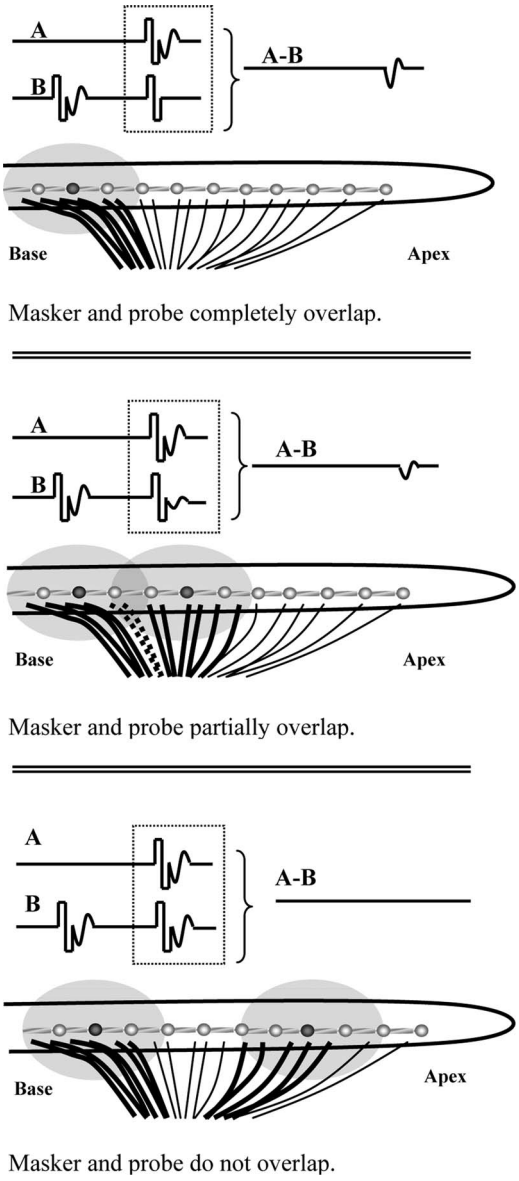


FIG. 1. Schematic illustrating the effect of changing masker electrode position on the subtracted electrically evoked compound action potential (ECAP) response. In the top panel, masker and probe are on the same electrode, which yields a maximum ECAP response in the subtracted trace (A-B). Gray halos represent hypothetical current fields. Nonbolded lines represent nonstimulated nerve fibers; bolded lines represent stimulated fibers. In the middle panel, masker and probe are on different electrodes (partially overlapping), which yields a smaller subtracted ECAP response. (Dashed nerve fibers represent those common to both the masker and probe electrodes.) In the bottom panel, masker and probe are on widely spaced electrodes, which results in no measurable ECAP in the subtracted response.

ECAP masking patterns for this study were obtained with the probe pulse fixed on one electrode while the location of the masker was varied across all other electrodes in the array. The recording electrode was generally fixed at a

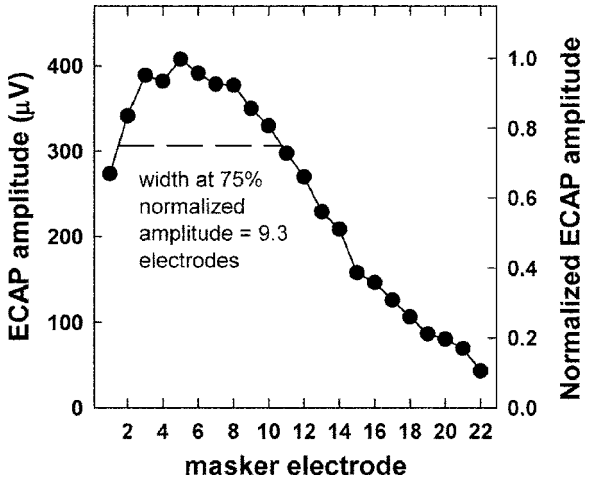


FIG. 2. Example of absolute (left ordinate) and normalized (right ordinate) ECAP amplitude plotted as a function of masker electrode for probe electrode 5. Data are from subject M35b. Dashed line represents width of the ECAP masking function at 75% normalized amplitude.

location two or three electrode positions apical to the probe. For some electrodes in some subjects, an ECAP could easily be recorded when the masker was placed on the same electrode that was used for recording. However, in most instances, a large stimulus artifact was measured in this condition. Consequently, the recording electrode was typically changed to either one position basal to or apical to its original position to allow for ECAP measurement with the masker on the original recording electrode. The recording electrode was then returned to its previous location when the masker was applied to subsequent electrodes in the array. This latter method was used for the majority of ECAP functions.

Masker electrodes were stimulated at 80% of the behavioral dynamic range (i.e., the 250 pps stimulus) for each electrode. The probe level was fixed at 80% of behavioral dynamic range for that electrode. This level was chosen to be consistent with presentation levels for the electrode pitch-ranking portion of this study. For subjects M24 and M50, the 80% level was inadequate for recording ECAP responses, so a 90% level was used. Levels for the psychophysical portion were likewise adjusted for these two subjects.

2. ECAP data analysis

ECAP waveforms were analyzed offline using a custom Matlab computer program. ECAP amplitudes were calculated from the first negative peak, N1, to the following positive peak or plateau. Figure 2 shows an example of ECAP amplitudes (left ordinate) plotted as a function of masker electrode for subject M35b with the probe pulse fixed on electrode 5. Typical response curves had largest amplitudes at or near the probe electrode, with decreasing amplitude as the distance between masker and probe electrodes increased.

ECAP amplitudes were normalized to the amplitude obtained with the masker and probe on the same electrode. Normalized ECAP amplitudes (right ordinate, Fig. 2) were then plotted as a function of each masker electrode. Alternatively, data could be plotted as a function of longitudinal distance (in millimeters) from the probe electrode rather than by electrode number since the distance between electrodes is different from base to apex within the Contour as well as between the Contour and straight arrays. However, the variable interelectrode distances within the Contour are designed as such to maintain the same radial distance from the modiolus as the straight array. If the electrode array in the Contour is positioned appropriately, then plotting these data as a function of electrode number should more accurately represent the distance between the electrode and neural population than longitudinal distance along the array. Thus, data in this study are plotted as a function of electrode number rather than millimeter distance from the probe electrode. In should be noted, however, that radiologic information was not available to confirm radial distance of the electrodes. The effect of device type on these experimental outcomes is examined further in a companion paper (Hughes and Abbas, 2006).

The width of the ECAP amplitude function was measured in number of electrodes at 75% of the normalized ECAP amplitude (horizontal dashed line in Fig. 2). The endpoints used to calculate width at 75% of the normalized amplitude were derived using the slope between adjacent electrodes' data points. The value of 75% was chosen for width measures because it was the lowest percent of normalized amplitude where a full width could successfully be measured for all tested electrodes in all subjects.

D. Electrode pitch ranking

1. Stimuli and procedure

The ability to discriminate between electrodes on the basis of pitch was tested using a two-interval, two-alternative forced choice (2I2AFC) pitch-ranking task. Using this method, two electrodes were sequentially stimulated and the subject responded as to whether the second sound was higher or lower in pitch relative to the first sound. This procedure was similar to that described by Laneau and Wouters (2004). Prior to testing, the subject was informed that the loudness of the sounds would vary slightly, but that they were to ignore loudness and only pay attention to the pitch. The subject responded verbally to the investigator, and this response was keyed into the computer by the investigator where it was automatically stored for later analysis. Neither visual cues nor feedback were provided to the subjects.

The stimulus for the electrode pitch-ranking task was generated by a custom computer program that utilized Nucleus Implant Communicator (NIC) subroutines (Cochlear Corp.). The stimulus was a 250-pps, 500-ms pulse train identical to that used to determine behavioral dynamic range. This stimulus was presented to a single intracochlear electrode (relative to MP1), followed by a 500-ms silent interval before the same stimulus was presented to a second electrode, also relative to MP1. Stimuli were randomly presented at 70%, 80%, and 90% of the dynamic range for each electrode to control for loudness cues. For the two subjects whose ECAP data were collected at 90% of the dynamic range, electrode pitch-ranking stimuli were presented at 80%, 90%, and 100% of the dynamic range.

Within a given trial, the first (standard) electrode stimulated was always the same electrode. This electrode served as a reference to which the pitch of the second electrode was

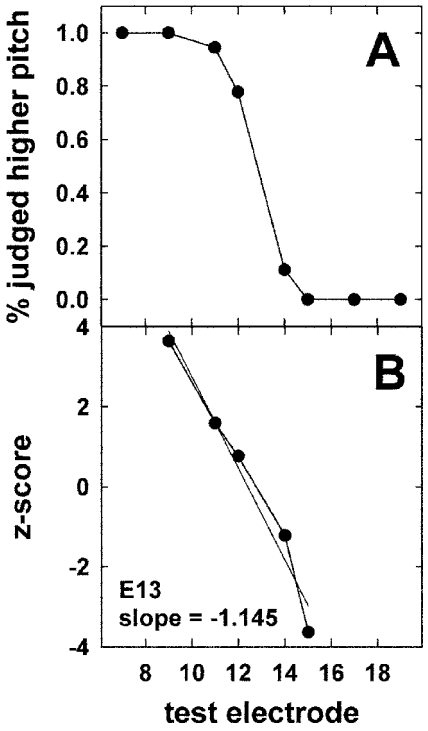


FIG. 3. Example showing methodology used to quantify results of electrode pitch ranking. Data are from standard electrode 13 in subject M35b. Panel (A) shows percent judged higher pitch for responses pooled across two trials. Panel (B) shows data from panel (A) converted to *z*-scores with data points beyond 0% and 100% omitted. Slope of linear regression analysis is indicated in panel (B).

compared. The location of the second (test) electrode was randomized and varied from one, two, four, or six electrode positions basal to and apical to the standard electrode, for a total of eight electrode pairs. The exception was for electrodes that were located less than six positions away from the basal or apical end or for short-circuit electrodes. For example, test electrodes for standard electrode 5 were electrodes 1, 2, 3, 4, 6, 7, 9, and 11. A total of 72 conditions were presented during one trial. These 72 conditions encompassed all combinations of electrode pairs and stimulus levels (one standard electrode at three roved levels multiplied by eight test electrodes at three roved levels each). Each of the eight electrode pairs were randomly presented nine times each, with each stimulus level randomized and presented an equal number of times. Each trial of 72 presentations was tested twice, but not sequentially, for a given standard electrode. Presentation order of each trial was randomized.

2. Data analysis

Results from the two trials for a given electrode pair were pooled and then quantified as a percent value for the number of times the subject responded that the second presentation was judged as higher in pitch relative to the first. These results were plotted as a function of test electrode. Figure 3(A) shows data from standard electrode 13 for Subject M35b. Percent values were expressed as a decimal (i.e., 100% = 1.0) and then converted to *z*-scores. For all subjects, once a score reached 100% or 0%, subsequent scores on more basal or apical electrodes were eliminated to avoid falsely reducing the slope of the psychometric function. For example, data points for test electrodes 7, 17, and 19 in subject M35b [Fig. 3(A)] were eliminated prior to converting percent scores to *z*-scores. In several cases, once the score reached 100% or 0%, scores on more distal test electrodes decreased or increased, respectively. Even in those cases subsequent scores beyond the initial 100% or 0% were eliminated so as to characterize the steepest portion of the function. *z*-scores served to normalize the percent data to a normal distribution. *z*-scores were then plotted as a function of test electrode for each standard electrode. Figure 3(B) shows data from Fig. 3(A) converted in this manner. Simple linear regression analysis was then applied to these data and the slope of the regression line was calculated. In the final step of data analysis, electrode pitch-ranking slope was plotted as a function of width of the ECAP function at 75% normalized amplitude for each subject.

III. RESULTS

A. Individual data

The primary goal of this study was to examine the relation between ECAP channel interaction and the ability to discriminate pitch between electrodes in a psychophysical pitch-ranking task. Figure 4 shows *z*-score data from the electrode pitch-ranking task (top row of graphs) and the cor-

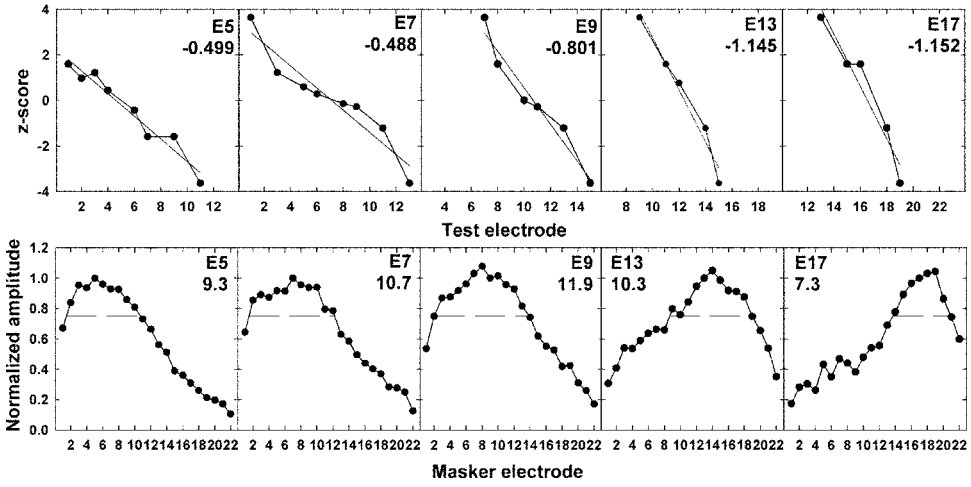


FIG. 4. Electrode pitch-ranking data (top row) and corresponding normalized ECAP masking functions (bottom row) for selected electrodes tested in subject M35b. Dashed line in the lower graphs represents width of the ECAP masking function at 75% normalized amplitude. Reference electrode and slope of regression line are indicated in the top row of the graphs; probe electrode and width at 75% normalized amplitude (in number of electrodes) are indicated in the bottom row of the graphs.

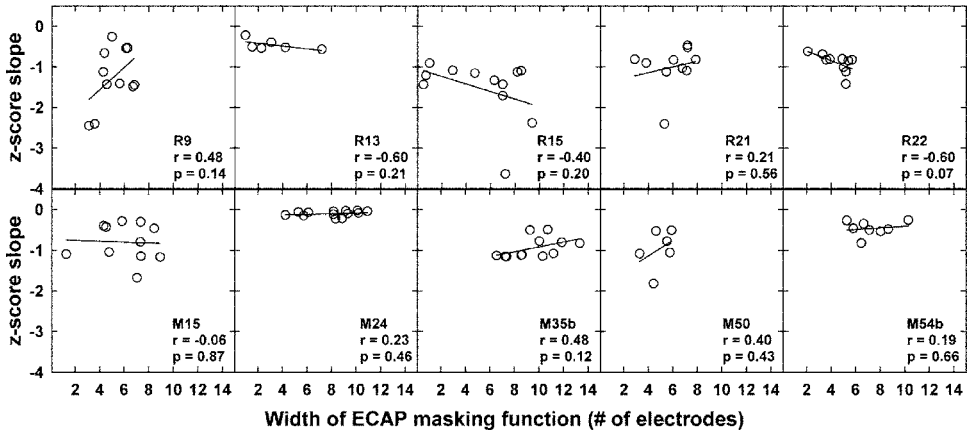


FIG. 5. Individual results showing width of the ECAP masking function versus slope of the electrode pitch-ranking function. Each panel represents data from a different subject (top row, 24R Contour array; bottom row, 24M straight array). Solid lines represent results of linear regression analyses. Subject number, correlation coefficient (r value), and p value are indicated on each graph.

responding normalized ECAP masking functions (bottom row of graphs) for a subset of electrodes tested for subject M35b. Standard electrode number and slope of the regression line are shown on each of the pitch-ranking graphs. Width of the ECAP masking function (in number of electrodes) at 75% normalized amplitude is indicated with probe electrode number on each of the ECAP graphs. For this subject, electrode pitch-ranking ability was poorest for the most basal electrodes, and best for electrodes in the apical half of the array. For some subjects, adjacent electrodes could be distinguished with 100% accuracy with monopolar stimulation. ECAP masking functions were sharpest for apical electrodes and broadest for middle electrodes in this subject. Various trends were noted across individuals for both measures.

To more directly evaluate the relation between electrode pitch-ranking ability and ECAP channel interaction, slope of the regression line from the z-score function was plotted relative to width of the ECAP function for each electrode tested within a subject. These results are shown in Fig. 5. Each graph represents data from a different subject and each symbol represents data from a different electrode. Subject numbers are indicated on each graph. The upper five graphs show data from subjects with the 24R Contour array and the lower five graphs show data from subjects with the 24M straight array. Linear regression lines are depicted in each graph. Pearson Product Moment Correlation was executed for each subject to evaluate the strength of association between width of the ECAP masking function and slope of the electrode pitch-ranking function. The resulting correlation coefficients and p values are shown in each graph. If these data were to support the hypothesis that less ECAP channel interaction is indicative of better electrode pitch discrimination ability, then we would expect to see linear regression lines for the plots in Fig. 5 to have a positive slope (i.e., narrow ECAP functions should correlate with steep slopes of the psychophysical functions).

None of the 10 subjects showed a significant positive correlation that was consistent with this hypothesis ($p > 0.05$ for all subjects). Three subjects (R9, M35b, and M50) showed a moderate positive correlation between ECAP masking functions and electrode pitch discrimination ability ($r=0.48$, $r=0.48$, and $r=0.40$, respectively). Three Contour subjects (R13, R15, and R22) showed a moderate correlation in the opposite direction ($r=-0.60$, $r=-0.40$, $r=-0.60$, respectively). There was no notable trend for the remaining four subjects (R21, M15, M24, and M54b). These results suggest that the width of the ECAP channel interaction function is not predictive of the ability to discriminate between electrodes on the basis of pitch.

B. Group data

Figure 6 shows the relation between width of the ECAP masking function and slope of the electrode pitch-ranking function for all electrodes in all 10 subjects. Filled and open circles represent data from subjects with the 24M straight array and 24R Contour array, respectively. Linear regression lines are plotted for each device type separately. Overall, there was no significant correlation between these two measures when data from all 10 subjects were pooled ($r=0.14$, $p=0.17$). Similarly, there was no significant correlation between the two measures when data from each device were analyzed separately ($r=0.13$, $p=0.35$, straight array; $r=-0.23$, $p=0.11$, Contour array). These results were consistent with individual results shown in Fig. 5. It should be noted, however, that there was a significant difference in width of the ECAP masking function between the two device types, with Contour subjects having significantly narrower widths

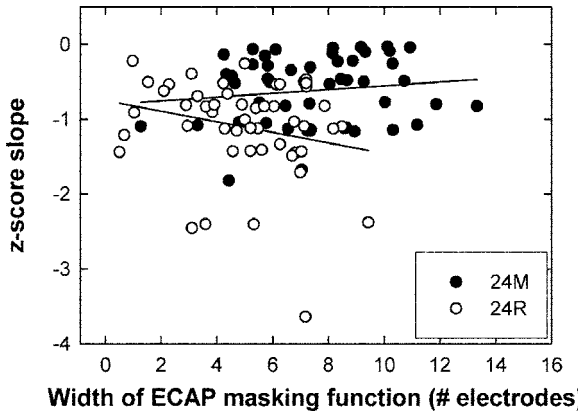


FIG. 6. Group data showing the relation between width of the ECAP masking function and slope of the electrode pitch-ranking function. Filled circles represent individual electrode data from subjects with the 24M straight array. Open circles represent individual electrode data from subjects with the 24R Contour array. Linear regression lines are plotted separately for each device type.

than straight array subjects ($p=0.007$). These results are discussed further in a companion paper (Hughes and Abbas, 2006).

C. Comparison to speech perception

A number of studies have evaluated the relation between electrode discrimination ability and speech perception for bipolar stimulation (Busby and Clark, 2000; Dawson *et al.*, 2000; Henry *et al.*, 2000; Nelson *et al.*, 1995; Throckmorton and Collins, 1999), common ground stimulation (Dawson *et al.*, 2000; Henry *et al.*, 2000), or pseudomonopolar stimulation (Dawson *et al.*, 2000). Townshend *et al.* (1987) reported electrode discrimination results for three subjects using monopolar stimulation, however these results were not compared to speech perception outcomes. There do not appear to be any published reports that have evaluated this relationship for true monopolar stimulation, which was the electrode configuration used in the present study.

Speech perception performance was compared to electrode pitch-ranking ability for three speech perception measures: HINT sentences in quiet, Consonant-Nucleus-Consonant (CNC) word lists (average of two lists) in quiet, and average percent correct phonemes for the CNC words. Individual speech perception scores for these subjects are reported in a companion paper (Hughes and Abbas, 2006). Because statistical analyses showed no effect of electrode place on electrode pitch-ranking outcomes ($p>0.05$), slope of the z-score function was averaged for all electrodes tested within a subject. Linear regression analysis was applied to all data (Contour and straight). Results showed no significant correlation between electrode pitch-ranking ability and performance on HINT sentences ($r=0.04$, $p=0.92$) when data from all subjects were analyzed together, or when data from each device type were analyzed separately ($r=0.19$, $p=0.76$, straight array; $r=-0.24$, $p=0.70$, Contour array). Similarly, there was no significant correlation between electrode pitch-ranking ability and performance on CNC words ($r=0.20$, $p=0.59$) or phonemes ($r=0.22$, $p=0.53$) when data from all subjects were analyzed together. There was also no relation between these measures when data from each device type were analyzed separately ($r=0.33$, $p=0.59$, straight array, words; $r=0.38$, $p=0.53$, straight array, phonemes; $r=-0.37$, $p=0.54$, Contour array, words; $r=-0.24$, $p=0.70$, Contour array, phonemes). Last, there was no significant difference in speech perception scores between the Contour and straight array for all three test measures ($p=0.71$ HINT, $p=0.06$ CNC words, $p=0.07$ CNC phonemes). These results do not support the hypothesis that better speech perception is related to better electrode pitch discrimination on the basis of pitch for monopolar stimulation.

Speech perception performance was also compared to the average width of the ECAP masking function for each subject. Results showed no statistically significant correlation between speech perception performance and average width of the ECAP masking function for any of the three speech perception measures ($r=0.24$, $p=0.50$, HINT; $r=0.33$, $p=0.35$, CNC words; $r=0.29$, $p=0.42$, phonemes). There was also no significant correlation when data were analyzed separately according to device type ($r=0.60$, $p=0.28$, straight array, sentences; $r=0.05$, $p=0.94$, straight array, words; $r=0.19$, $p=0.76$, straight array, phonemes; $r=-0.49$, $p=0.41$, Contour array, sentences, $r=-0.30$, $p=0.63$, Contour array, words; $r=-0.47$, $p=0.42$, Contour array, phonemes).

IV. DISCUSSION

The primary goal of this study was to examine the relation between ECAP channel interaction and the ability to discriminate pitch between electrodes in a psychophysical pitch-ranking task. It was hypothesized that subjects should perform better on an electrode pitch-ranking task when using electrodes with less channel interaction as measured with the ECAP. Results from this study showed no clear evidence that the width of ECAP masking functions are related to electrode pitch discrimination ability for individual subjects or for group data.

A. Quantifying ECAP masking patterns

One possible reason for the lack of correlation between width of the ECAP masking function and electrode pitch discrimination ability in this study may be the method used to quantify the ECAP masking patterns. It is possible that a simple measure of width is not the best way to quantify the ECAP patterns. McKay *et al.* (1999) suggested that changes in pitch percept with broad modes of stimulation might be due to shifts in the edges or peaks of the regions of excitation. Pfingst *et al.* (1997) suggested that pitch discrimination with broad modes of stimulation (i.e., monopolar) is achieved with more neurons operating at lower discharge rates. Thus, a simple measure of width at a 75% reduction in amplitude does not adequately characterize other aspects of the ECAP masking function that may be important to consider when comparing such measures to electrode pitch discrimination ability. Such aspects include differences in location of the peak of the function relative to the probe electrode, shape or elevation of the function at more distal masker electrode locations (i.e., overall amount of masking), or symmetry (or asymmetry) about the probe electrode.

1. Location of maximum masking

Based on previous studies, we would expect that the peak of the ECAP masking function would occur when the masker and probe are on the same electrode, indicating maximum amount of masking (Abbas *et al.*, 2004; Cohen *et al.*, 2003). However, the peak of the ECAP masking function can occur for masker locations distal to the probe electrode in some cases. For example, the peak of excitation in the ECAP masking function for probe electrode 17 may actually occur at masker electrode 19 (see Fig. 4, E13 and E17 as examples). This trend was exhibited for several electrodes in subjects participating in this study. In addition, Abbas *et al.* (2004) showed similar examples using this same measure in other cochlear implant users, and other investigators have shown these examples with psychophysical forward masking patterns (Chatterjee and Shannon, 1998; Throckmorton and Collins, 1999). The most likely explanation for

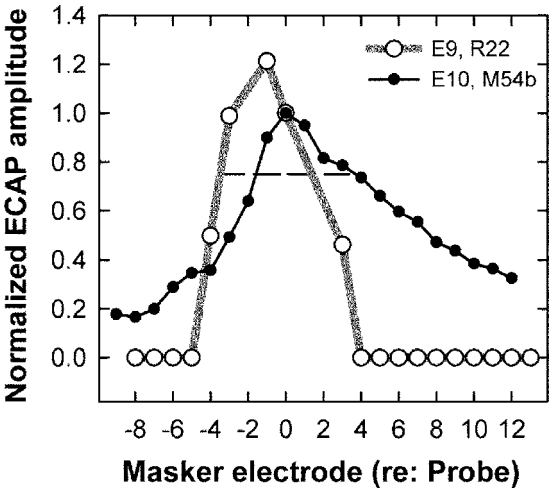


FIG. 7. Normalized ECAP masking functions for two different subjects with similar widths but different overall masking patterns. Dashed horizontal line represents width at 75% normalized amplitude.

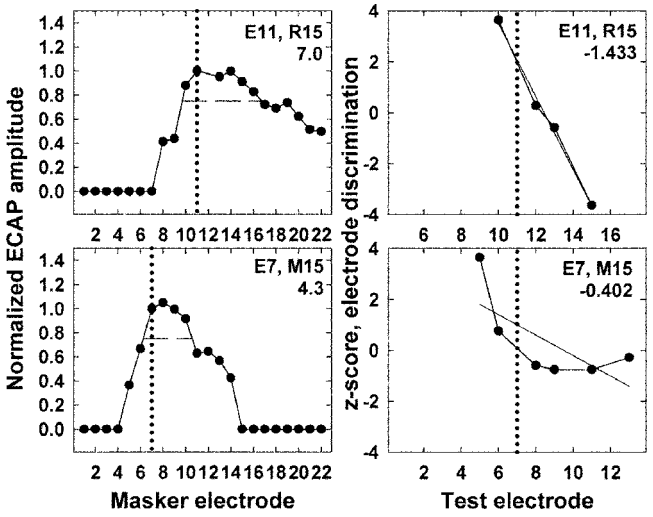


FIG. 8. Individual examples of asymmetric ECAP functions (left-hand column) and corresponding electrode pitch-ranking functions (right-hand column). Electrode number and subject number are indicated on each graph.

the occurrence of a shift in the peak of excitation is uneven patterns of neural survival, or "dead" regions in the vicinity of the probe electrode (Abbas *et al.*, 2004). We attempted to account for shifted peaks in this study by normalizing to the masker-equals-probe electrode rather than to the electrode with the maximum amount of masking. If ECAP functions were normalized in the latter manner, then widths would appear to be narrower than they actually are, and the width measure would not be sensitive to conditions where peaks are displaced distal to the probe electrode.

2. Total amount of masking across the array

The width of the ECAP function at 75% normalized amplitude does not quantify the overall shape or elevation of the function at masker electrodes far away from the probe. For example, two subjects might have similar width measures for the ECAP masking function at 75% normalized amplitude, but their overall shapes or total amount of interaction may be very different. For some subjects, ECAP amplitude measures drop quickly to zero at a few masker electrode locations beyond 75% normalized amplitude. For other subjects, the overall function may be rather broad and normalized amplitude for maskers located on the extreme basal or apical ends of the function may never reach zero. Figure 7 shows an example of two ECAP masking functions with similar width measures (~5 electrodes; depicted by the horizontal dashed line) but very different overall masking patterns. The ECAP masking function for probe electrode 9 in Subject R22 (open circles) has a width of 4.9 electrodes with amplitude measures of zero on a total of 14 masker electrodes located distal to the probe. Thus, this function shows a relatively distinct stimulation pattern and no overlap with distant neural populations. The function also shows a displaced peak at masker electrode 8 (probe is electrode 9). In contrast, the ECAP masking function for probe electrode 10 in subject M54b (filled circles) has a width of 5.3 electrodes, however ECAP amplitude never reaches zero for any of the masker electrodes. Additionally, the peak of this function occurs at the probe electrode. This comparison shows that despite having similar width measures at 75% normalized amplitude, more focused stimulation patterns exist for some subjects (i.e., subject R22), while other subjects show that some amount of interaction remains for the most widely spaced masker and probe electrodes (i.e., subject M54b). It is also interesting to note that subject R22 had a much steeper pitch-ranking function (slope=−0.81) than subject M54b (slope=−0.27) for those respective electrodes. It is possible that the width of the entire masking function (i.e., width where the function reaches zero) is more important than the width at a given reduction in amplitude. Unfortunately, it was not possible to measure the entire width of the masking function for many electrodes in the present study because not all functions reached zero. It was also not possible to measure widths at lower percentages of normalized amplitude (e.g., 25% or 50%) because many functions did not decrease to these lower levels toward the ends of the array. Widths for this study were measured at 75% because that was the lowest percent of normalized amplitude where a full width could successfully be measured for all tested electrodes in all subjects.

3. Asymmetry of functions

We might expect that the shape of the ECAP masking function would be symmetric about the probe electrode, however in actuality, the shape of the function may be very different for maskers located basal to the probe versus maskers located more apically. For example, in the top left-hand graph of Fig. 8, the ECAP masking function for probe electrode 11 (subject R15) drops steeply as the masker is applied to electrodes in a more basal direction relative to the probe (vertical dotted line represents probe electrode location). When the masker is just four electrode positions away (on E7), the ECAP amplitude has already reached zero. However, masker electrodes applied to more apical electrodes show a shallow decrease in amplitude as the distance from the probe electrode increases. For masker electrode 22, the normalized amplitude drops to just about 50%. This pat-

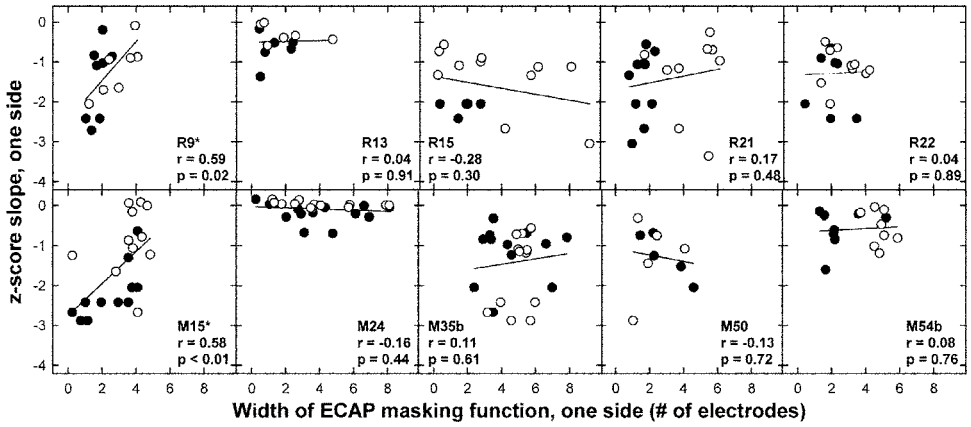


FIG. 9. Individual data from Fig. 5 replotted so that basal (filled circles) and apical (open circles) halves of each function are represented separately. Subject number, correlation coefficient (r value), and p value are indicated on each graph. Asterisks denote significant correlations.

tern shows that there is a greater amount of masking for electrodes located apical to the probe for electrode 11 in this subject. These results concur with those reported by Cohen *et al.* (2003). A similar asymmetrical pattern is observed for electrode 7 in subject M15 (Fig. 8, bottom left-hand panel). The asymmetries seen in various ECAP masking functions may be due to uneven patterns of nerve survival and/or nonuniform current spread throughout the cochlea, particularly since the cochlea is tapered from base to apex. Thus, a width measure at 75% of the normalized amplitude does not fully characterize these asymmetries.

B. Quantifying electrode pitch ranking

Asymmetry of functions: Like the ECAP masking functions, a number of the psychophysical z-score functions were asymmetric about the standard electrode. The right-hand panels in Fig. 8 show corresponding z-score functions for the ECAP masking patterns discussed above. These examples show much better electrode pitch-ranking ability for basal-side electrodes versus apical-side electrodes. The vertical dotted line in each graph highlights the standard electrode location. For subject R15, perfect discrimination was achieved at one electrode position in a basal direction; however, apical comparisons required an electrode separation of four positions to achieve perfect discrimination. Similarly, subject M15 reached 100% for two electrode positions in a basal direction, whereas chance levels of performance (i.e., z-score of zero is 50%) were achieved for electrodes located as far as six positions in an apical direction. Thus, overall slope of the linear regression line does not adequately capture the asymmetries in electrode pitch-ranking ability for test electrodes located in an apical versus basal direction relative to the standard electrode. It is also worthwhile to note some examples that show the opposite trend. For a subset of electrodes in some subjects, poorer electrode pitch ranking occurred on the basal side where the ECAP function was steeper. As a result of these findings, it was of interest to characterize both the ECAP and psychophysical functions separately for each side of the function (i.e., basal side and apical side) to better quantify these differences.

Figure 9 shows data from Fig. 5 plotted separately for the basal half (filled circles) and apical half (open circles) of each function. Width of the ECAP masking function was calculated from the probe electrode to the point where the function dropped to 75% of the normalized amplitude basal to the probe (filled circles) or apical to the probe (open circles). These measures were compared to the slope of the z-score function, calculated for test electrodes located either basal to or apical to the standard electrode. When data were analyzed in this manner, a significant positive correlation between half-width of the ECAP masking function and half-slope of the z-score function was found for subject R9 ($r=0.59$, $p=0.02$) and subject M15 ($r=0.58$, $p<0.01$). No significant correlation was noted for the remaining eight subjects. Correlation coefficients and p values are indicated along with subject number in each graph in Fig. 9 (asterisks indicate statistical significance for $\alpha=0.05$). In general, these results are similar to those reported for Fig. 5.

C. Interaction versus pitch ranking

The general lack of correlation between ECAP channel interaction functions and electrode pitch-ranking ability for individuals in this study may be due to the dissimilarity between the two tasks and what they measured. The ECAP masking functions in this study employed a forward masking technique with single pulses that assessed spatial and temporal interactions between neural populations. Those measures depend upon the spread of excitation to both the masker and the probe and the relative location of the recording electrode, and are completely peripheral measures. The electrode pitch-ranking task used a different stimulus (pulse trains) and assessed spatial/spectral resolution. That task, being a subjective measure, also required a large component of central processing mechanisms, including the ability to separate loudness variations from pitch variations. It may be that the amount of neural interaction measured with the ECAP was not a sensitive measure of the peripheral activity that contributes to electrode pitch-ranking ability. In other words, electrode pitch-ranking ability may be more closely related to shifts in the edges of the peripheral excitation patterns, relative location of the peak of excitation, or relative firing rates of individual fibers (i.e., McKay *et al.* 1999; Pfingst *et al.*, 1997).

Cohen *et al.* (2003) reported a significant relation between width of the ECAP masking function and width of the psychophysical forward masking function for increased distance between masker and probe electrodes. Additionally, Throckmorton and Collins (1999) reported a significant rela-

tion between psychophysical forward masking functions obtained with spatial separation between masker and probe versus an electrode discrimination task. Thus, it could be postulated that the ECAP masking width would be correlated with electrode discrimination ability. However, the present study did not assess electrode discrimination in the same way as the Throckmorton and Collins study. That study used loudness-balanced stimuli in a 2IFC procedure, where the subjects reported whether the two stimuli were the same or different. That type of task allows the subject to make the same/different determination on either the basis of pitch or any other perceptual difference (Collins *et al.*, 1997; Laneau and Wouters, 2004). In the present study, electrode discrimination was tested on the basis of pitch only, where subjects responded as to whether the second stimulus was higher or lower in pitch relative to the first. Thus, it may be that ECAP masking functions are more closely related to an electrode discrimination task where discrimination is based on any perceptual difference between electrodes rather than on pitch alone.

Another possible explanation for the lack of correlation between ECAP channel interaction functions and electrode pitch-ranking ability for individuals in this study may be that central auditory neural pathways simply process peripheral stimulation differently across electrodes and/or subjects. Thus, the relation between pitch perception and ECAP masking patterns may differ between basal versus apical electrodes or from subject to subject. Last, it is possible that there is a relation between the amount of interaction among electrodes (measured physiologically) and electrode pitch-ranking ability within subjects, but the effect is too small to be measured with the methods used in this study.

D. Relation to speech perception

A number of investigators have reported a significant relation between electrode discrimination ability and speech perception (i.e., Busby and Clark, 2000; Henry *et al.*, 2000; Nelson *et al.*, 1995; Throckmorton and Collins, 1999). Specifically, better electrode discrimination ability (or a greater number of distinct pitch percepts) has been shown to correlate with higher speech perception scores for various speech tests. Results from the present study do not concur with these previous reports and are consistent with outcomes reported by Zwolan *et al.* (1997). There were two significant differences between the present study and these previous studies: first, all of the cited reports used bipolar stimulation while the present study used monopolar stimulation. Second, the majority of those studies (the exception being Nelson *et al.*, 1995) used an electrode discrimination task in which the subject had to choose which sound was different, rather than discriminate between electrodes solely on the basis of pitch (i.e., pitch ranking).

Another difference is in the specific tests that were used to measure speech perception. For example, Throckmorton and Collins (1999) showed significant relations between electrode discrimination and vowel, consonant, phoneme, and sentence recognition, but not word recognition. The present study did not assess vowel and consonant recognition separately. However, like Throckmorton and Collins, the present study showed no significant correlation between electrode discrimination ability and word recognition. It may be that vowel/consonant tests would have been a more sensitive measure to use in the present study. There also appeared to be significant ceiling effects for HINT sentences in the present study.

Another explanation may simply have to do with speech processor settings. The speech perception measures used in this study were made with the subject's everyday, most-used program and settings (all subjects used SPEAK or ACE). Therefore, processing strategies, processor type (i.e., body worn versus behind-the-ear), volume/sensitivity, and stimulation rates varied across subjects. It is possible that the subject's most-used program was not necessarily optimal for them. Nelson *et al.* (1995) reported excellent electrode discrimination ability for some subjects with poorer speech recognition ability. They suggested that the current speech processing strategy (probably F0/F2, although not explicitly stated in that study) was not sufficient to code speech in a manner that would take advantage of the excellent electrode discrimination ability. Although advances in speech coding strategies have been made since the Nelson *et al.* study was presented, the same argument could potentially be made here.

Last, the method used to quantify electrode discrimination ability differed across studies (e.g., d' measures, difference limens, slope of z-score function). This is possibly a factor contributing to different findings across studies, given the results from Collins *et al.* (1997), which showed different outcomes between electrode discrimination measures and estimates of electrode discrimination ability based on a pitch scaling procedure for the same subjects.

V. CONCLUSIONS

The primary hypothesis for this study was that electrodes with less channel interaction would have more distinct or accurate pitch percepts. In summary, the present study showed no relation between electrode pitch discrimination ability and electrophysiologic measures of channel interaction for individual subjects or for group data. The lack of correlation suggests that the amount of neural interaction measured with the ECAP was not indicative of differences in pitch perception between electrodes. This outcome may be attributed to methods used to quantify the data, as well as central processing components involved in the discrimination task.

This study showed that adjacent electrodes could be distinguished when using monopolar stimulation, which is consistent with reports for narrow bipolar stimulation (i.e., Busby and Clark, 2000; Nelson *et al.*, 1995). In addition, there was no significant relation between speech perception performance and either electrode pitch-ranking ability or width of the ECAP masking functions. Although radiologic data were not available for subjects in the present study, it may be valuable to obtain imaging studies to confirm elec-

trode position within the cochlea to determine how radial distance is related to ECAP masking patterns and/or electrode pitch discrimination ability.

ACKNOWLEDGMENTS

This study is based on doctoral dissertation work at the University of Iowa by the first author. The second author served as adviser and chair of the dissertation committee. This study was supported by funds from the NIH/NIDCD (Grant No. P50 DC00242), NIH/NCRR (Grant No. RR00059). The authors would like to thank George Woodworth for assistance with statistical analyses, and Arik Wald for programming support. Additional thanks go to the 10 subjects who participated in this study.

Abbas, P. J., and Brown, C. J. (**1988**). "Electrically evoked brainstem potentials in cochlear implant patients with multi-electrode stimulation," Hear. Res. **36**, 153–162.

Abbas, P. J., Brown, C. J., Shallop, J. K., Firszt, J. B., Hughes, M. L., Hong, S. H., and Staller, S. J. (**1999**). "Summary of results using the Nucleus CI24M implant to record the electrically evoked compound action potential," Ear Hear. **20**, 45–59.

Abbas, P. J., Hughes, M. L., Brown, C. J., Miller, C. A., and South, H. (**2004**). "Channel interaction in cochlear implant users evaluated using the electrically evoked compound action potential," Audiol. Neuro-Otol. **9**, 203–213.

Boex, C., Kos, M. -I., and Pelizzone, M. (**2003**). "Forward masking in different cochlear implant systems," J. Acoust. Soc. Am. **114**, 2058–2065.

Brown, C. J., Abbas, P. J., Borland, J., and Bertschy, M. R. (**1996**). "Electrically evoked whole nerve action potentials in Ineraid cochlear implant users: Responses to different stimulating electrode configurations and comparison to psychophysical responses," J. Speech Hear. Res. **39**, 453–467.

Brown, C. J., Abbas, P. J., and Gantz, B. J. (**1990**). "Electrically evoked whole-nerve action potentials: Data from human cochlear implant users," J. Acoust. Soc. Am. **88**, 1385–1391.

Brown, C. J., Abbas, P. J., and Gantz, B. J. (**1998**). "Preliminary experience with Neural Response Telemetry in the Nucleus CI24M cochlear implant," Am. J. Otol. **19**, 320–327.

Brown, C. J., Hughes, M. L., Luk, B., Abbas, P. J., Wolaver, A. A., and Gervais, J. P. (**2000**). "The relationship between EAP and EABR thresholds and levels used to program the Nucleus CI24M speech processor: Data from adults," Ear Hear. **21**, 151–163.

Busby, P. A., and Clark, G. M. (**2000**). "Electrode discrimination by early-deafened subjects using the Cochlear Limited multiple-electrode cochlear implant," Ear Hear. **21**, 291–304.

Cartee, L. A., van den Honert, C., Finley, C. C., and Miller, R. L. (**2000**). "Evaluation of a model of the cochlear neural membrane. I. Physiological measurement of membrane characteristics in response to intrameatal electrical stimulation," Hear. Res. **146**, 143–152.

Chatterjee, M., and Shannon, R. V. (**1998**). "Forward masked excitation patterns in multielectrode electrical stimulation," J. Acoust. Soc. Am. **103**, 2565–2572.

Cohen, L. T., Richardson, L. M., Saunders, E., and Cowan, R. S. C. (**2003**). "Spatial spread of neural excitation in cochlear implant recipients: comparison of improved ECAP method and psychophysical forward masking," Hear. Res. **179**, 72–87.

Collins, L. M., Zwolan, T. A., and Wakefield, G. H. (**1997**). "Comparison of electrode discrimination, pitch ranking, and pitch scaling data in postlingually deafened adult cochlear implant subjects," J. Acoust. Soc. Am. **101**, 440–455.

Dawson, P. W., McKay, C. M., Busby, P. A., Grayden, D. B., and Clark, G. M. (**2000**). "Electrode discrimination and speech perception in young children using cochlear implants," Ear Hear. **21**, 597–607.

Dillier, N., Lai, W. K., Almqvist, B., Frohne, C., Muller-Deile, J., Stecker, M., and von Wallenberg, E. (**2002**). "Measurement of the electrically evoked compound action potential via a neural resonse telemetry system," Ann. Otol. Rhinol. Laryngol. **111**, 407–414.

Gantz, B. J., Tyler, R. S., Knutson, J. F., Woodworth, G., Abbas, P., McCabe, B. F., Hinrichs, J., Tye-Murray, N., Lansing, C., Kuk, F., and Brown, C. (**1988**). "Evaluation of five different cochlear implant designs: Audiologic assessment and predictors of performance," Laryngoscope **98**, 1100–1106.

Henry, B. A., McKay, C. M., McDermott, H. J., and Clark, G. M. (**2000**). "The relationship between speech perception and electrode discrimination in cochlear implantees," J. Acoust. Soc. Am. **108**, 1269–1280.

Hughes, M. L., and Abbas, P. J. (**2006**). "Electrophysiologic channel interaction, electrode pitch ranking, and behavioral threshold in straight versus perimodiolar cochlear implant arrays." J. Acoust. Soc. Am. **119**, 1538–1548.

Laneau, J., and Wouters, J. (**2004**). "Multichannel place pitch sensitivity in cochlear implant recipients," J. Assoc. Res. Otolaryngol. **5**, 285–294.

McKay, C. M., O'Brien, A., and James, C. J. (**1999**). "Effect of current level on electrode discrimination in electrical stimulation," Hear. Res. **136**, 159–164.

Nelson, D. A., Van Tasell, D. J., Schroder, A. C., Soli, S., and Levine, S. (**1995**). "Electrode ranking of "place pitch" and speech recognition in electrical hearing," J. Acoust. Soc. Am. **98**, 1987–1999.

Peeters, S., Van Immerseel, L., Zarowski, A., Houben, V., Govaerts, P., and Offeciers, E. (**1998**). "New developments in cochlear implants," Acta Otorhinolaryngol. Belg. **52**, 115–127.

Pfingst, B. E., Zwolan, T. A., and Holloway, L. A. (**1997**). "Effects of stimulus configuration on psychophysical operating levels and on speech recognition with cochlear implants," Hear. Res. **112**, 247–260.

Shannon, R. V. (**1983**). "Multichannel electrical stimulation of the auditory nerve in man. II. Channel interaction," Hear. Res. **12**, 1–16.

Shannon, R. V. (**1985**). "Loudness summation as a measure of channel interaction in a cochlear prosthesis," in *Cochlear Implants*, edited by R. A. Schindler and M. M. Merzenich (Raven Press, New York), pp. 323–334.

Throckmorton, C. S., and Collins, L. M. (**1999**). "Investigation of the effects of temporal and spatial interactions on speech-recognition skills in cochlear-implant subjects," J. Acoust. Soc. Am. **105**, 861–873.

Townshend, B., Cotter, N., Van Compernolle, D., and White, R. L. (**1987**). "Pitch perception by cochlear implant subjects," J. Acoust. Soc. Am. **82**, 106–115.

White, M. W., Merzenich, M. M., and Gardi, J. N. (**1984**). "Multichannel cochlear implants: Channel interaction and processor design," Acta Oto-Laryngol. **110**, 493–501.

Zwolan, T. A., Collins, L. M., and Wakefield, G. H. (**1997**). "Electrode discrimination and speech recognition in postlingually deafened adult cochlear implant subjects," J. Acoust. Soc. Am. **102**, 3673–3685.

Electrophysiologic channel interaction, electrode pitch ranking, and behavioral threshold in straight versus perimodiolar cochlear implant electrode arrays[a)]

Michelle L. Hughes[b)] and Paul J. Abbas
Department of Otolaryngology—Head and Neck Surgery, University of Iowa, 200 Hawkins Drive, Iowa City, Iowa 52242; Department of Speech Pathology and Audiology, University of Iowa, Iowa City, Iowa 52242

(Received 31 May 2005; revised 9 December 2005; accepted 12 December 2005)

The primary goal of this study was to examine electrophysiologic measures of channel interaction, electrode pitch discrimination ability using a pitch-ranking task, and behavioral threshold levels in individuals implanted with a straight electrode array versus a perimodiolar array. It was hypothesized that perimodiolar arrays should yield lower thresholds, less channel interaction as measured with the electrically evoked compound action potential (ECAP), and better electrode pitch-ranking ability. Results from ten adult Nucleus 24 recipients ($N=5$ straight array, $N=5$ perimodiolar Contour array) showed no significant difference in threshold between the two electrode designs; however, there was significantly better electrode pitch-ranking ability and less channel interaction as measured with the ECAP for perimodiolar electrodes. Additionally, there was a significant positive correlation between behavioral threshold and width of the ECAP interaction function for Contour group data. There was no significant correlation between behavioral threshold and electrode pitch-ranking ability. These data suggest that electrode design and/or perimodiolar position may reduce physiologic channel interaction in the cochlea and improve electrode pitch discrimination ability; however, this positive finding did not translate into significantly better speech perception ability for Contour subjects.
 [DOI: 10.1121/1.2164969]

PACS number(s): 43.64.Me, 43.64.Pg, 43.66.Fe, 43.66.Cb [BLM] Pages: 1538–1547

I. INTRODUCTION

Recent cochlear implant designs have focused on placing the electrode array closer to the modiolus in an attempt to achieve lower thresholds and reduce excess current spread that may result in channel interaction. Previous research has described current field patterns for various electrode configurations in the electrically stimulated cochlea. The results of these experiments show that the spatial gradient of the electric field increases as the distance between the stimulating electrodes and neural elements decreases (Briaire and Frijns, 2000; Finley *et al.*, 1990; Frijns *et al.*, 1995; Kral *et al.*, 1998; Ranck, 1975; Rattay *et al.*, 2001). In other words, at closer electrode-nerve distances there is a steeper slope of the electric field, whereas at farther distances the slope of the electric field is shallower. Therefore, an electrode that is closer to surviving neural elements would need to inject less current for the population to reach discharge threshold due to the steeper slope of the electric field. The result would be lower thresholds for proximally placed electrode arrays. Steeper electric fields yield less current spread to adjacent neural populations, which would also decrease the likelihood of channel interaction (Frijns *et al.*, 1996). Less channel interaction in the cochlea may result in better pitch discrimination between electrodes due to minimized overlap of current fields or greater differences in shifts in the edges of excitation patterns between electrodes with more localized current fields (McKay *et al.*, 1999; Pfingst *et al.*, 1999). The primary purpose of this study was to determine whether perimodiolar electrode placements yield lower behavioral threshold, less electrophysiologic channel interaction, and better electrode pitch discrimination ability compared with traditional lateral-wall electrode arrays.

A number of studies have examined the effects of electrode proximity on behavioral threshold. Kawano *et al.* (1998) compared intracochlear factors in postmortem Nucleus 22 users with premortem psychophysical measures and found that behavioral thresholds (T levels) and maximum comfort levels (C levels) were positively correlated with the distance between electrodes and Rosenthal's canal; that is, lower T- and C levels were found in subjects whose electrode arrays were closer to the modiolus. Recent studies comparing perimodiolar electrode position to devices with traditional lateral wall placements have shown lower thresh-

[a)]Portions of this work were presented in "ECAP channel interaction and electrode discrimination ability in cochlear implants," Proceedings of VIII International Cochlear Implant Conference, Indianapolis, IN, May 2004; "The relationship between electrophysiologic and psychophysical measures of spatial spread in cochlear implants," Proceedings of Scientific and Technology Meeting of the American Auditory Society, Scottsdale, AZ, March 2003; 26th Midwinter Meeting of the Association for Research in Otolaryngology, Daytona Beach, FL, February 2003; and 7th International Cochlear Implant Conference, Manchester, United Kingdom, September 2002.

[b)]Present affiliation: Boys Town National Research Hospital, Lied Learning and Technology Center, 425 North 30th St., Omaha, NE 68131. Electronic mail: hughesml@boystown.org

olds and comfortable loudness levels with the perimodiolar design (Kreft *et al.*, 2004; Parkinson *et al.*, 2002; Saunders *et al.*, 2002; Young and Grohne, 2001).

In addition to examining the relation between behavioral levels and modiolus proximity, Cohen *et al.* (2001) also measured electrode discrimination ability in three subjects implanted with a precurved electrode array. In two of the three subjects, a portion of the electrode array occupied a space close to the modiolus. Those subjects had smaller pitch difference limens for electrodes that were closer to the modiolus in contrast to other electrodes in the array that occupied a more lateral position. These results suggest better spatial selectivity for electrodes positioned closer to neural elements.

Proximal location of the stimulating electrode to the neural population also seems to have implications for both lower physiologic thresholds and better spatial selectivity. Shepherd *et al.* (1993) reported lower electrically evoked auditory brainstem response (EABR) thresholds and reduced slope of the amplitude growth function for electrodes positioned adjacent to the modiolus and spiral ganglion or below the osseus spiral lamina versus electrodes placed near the lateral wall or in the middle of the scala tympani. Liang *et al.* (1999) showed that lower single-unit thresholds were related to better spatial selectivity for electrically stimulated auditory-nerve fibers in cats. They reported large variations in how different nerve fibers responded to changes in electrode location, and suggested that the more selective nerve fibers were closer to the stimulating electrode. Last, Cohen *et al.* (2003) reported narrower widths of masking functions obtained with the electrically evoked compound action potential (ECAP) for subjects implanted with the perimodiolar Nucleus Contour array and wider functions for subjects implanted with the Nucleus 24 M straight array. These results suggest less physiologic channel interaction for the perimodiolar electrodes.

The selectivity of electric fields on neural stimulation depends upon a number of factors in addition to electrode-nerve proximity. These other factors include nerve survival patterns, variability of stochastic characteristics of individual nerve fibers, cochlear anatomy, signal characteristics, and electrode design characteristics. Many of these factors can be interdependent and can affect threshold as well as selectivity. For example, an individual with poor nerve survival should have poorer selectivity and higher thresholds, presumably due to larger current requirements to achieve neural excitation. Additionally, a full-banded electrode array (such as the Nucleus 24M straight array) may produce less selective current field patterns compared to a half-band electrode array oriented toward the modiolus (such as the Nucleus 24R Contour array).

The primary goal of this study was to compare *T* level, electrophysiologic measures of channel interaction using the ECAP, and psychophysical electrode pitch discrimination ability in subjects implanted with a perimodiolar array versus a straight array. It is hypothesized that perimodiolar electrodes should have lower *T* levels, more spatially selective physiologic responses, and better electrode pitch discrimination ability compared to electrode arrays with lateral wall placements. A secondary goal of this study was to examine the relation between *T* level and electrophysiologic measures of spatial selectivity as well as the relation between *T* level and psychophysical measures of electrode pitch discrimination within subjects and between device types. Although lower *T* levels can result from either better neural survival or electrode proximity to neural elements, there is not a simple method to parse out these two effects. However, we would expect that lower *T* levels should be correlated with more discrete stimulation patterns within the cochlea because either the effective current fields are steeper and narrower (for more proximal electrode placements) or because there is a denser neural population. It was hypothesized that subjects should perform better on an electrode pitch-ranking task when using electrodes with lower *T* levels, and electrodes with lower *T* levels should exhibit less channel interaction as measured with the ECAP.

II. METHODS

A. Subjects

Ten adult Nucleus 24 cochlear implant recipients participated in this study. Subjects were divided into two groups. The straight-array group consisted of five subjects implanted with the Nucleus 24M device, and the perimodiolar group consisted of five subjects implanted with the Nucleus 24R(CS) Contour array. Subjects were primarily recruited on the basis of having large differences in *T* level across electrodes to maximize the ability to quantify any trends that might be seen across electrodes within a subject as well as across subjects. If *T* level were indeed indicative of either modiolar proximity or neural population density, then we would expect to see differences in physiologic spatial selectivity and/or differences in electrode pitch discrimination across electrodes for individuals with large differences in *T* level across electrodes. However, this criterion was difficult to meet for many 24M recipients. Demographic data for participants are listed in Table I. Subjects M35b and M54b each received bilateral implants within a single surgery; however, only one ear was tested for each subject. For bilateral subjects, the choice of test ear was also based on differences in *T* level across electrodes. Subject R9 had initially been implanted with an experimental "short electrode" (Nucleus CI 6+16+2M) cochlear implant for 9 months prior to receiving the Contour array in the same ear. Subject R22 presented with short circuits on electrodes 10 and 11, and subject M50 reported nonauditory percepts on electrodes 12 and 13 (which had previously presented as short circuits), so those electrodes were not tested in this study. Subject M50 also had four stiffening rings outside of the cochlea as noted at the time of surgery, indicating a shallower insertion depth.

B. Behavioral *T* levels and *C* levels

1. Equipment and stimuli

T levels and *C* levels were measured using the WinDPS clinical programming software (R116, build 445) installed on a desktop computer, which controlled a SPrint body-worn speech processor through the processor control interface (PCI). Stimuli consisted of a train of 25-μs/phase biphasic current pulses presented at a rate of 250 pps. Pulse train

TABLE I. Demographic data for subjects participating in this study.

Subject number	Device type	Ear	Age at implant	Duration implant use	Duration deafness	Reported etiology of deafness	Difference in *T* levels[a]	Electrodes tested
M15	24M	Right	69 yrs	5 yrs	10 yrs	Infection	7	5,6,7,9,11–17
M24	24M	Left	51 yrs	4.5 yrs	48 yrs	Trauma	38	6–18
M35b[b]	24M	Left	69 yrs	4 yrs	1 yr	Noise	6	5–7,9–11,13–18
M50	24M	Left	72 yrs	3 yrs	5 yrs	Sudden	30	5,8,9,11,16,19
M54b[b]	24M	Right	52 yrs	3 yrs	7 yrs	Unknown	16	3–6,8,10,11,13,15,17
R9	24R	Right	35 yrs	2 yrs	2 yrs	Unknown	15	3–5,7,8,10,11,13–15,17
R13	24R	Right	66 yrs	2 yrs	5 yrs	Prog/hered	16	5,7,10,13,15,18
R15	24R	Right	39 yrs	2 yrs	3 yrs	Hereditary	70	4–6,8,10–16,18
R21	24R	Right	68 yrs	5 mos	10 yrs	Unknown	29	4,5,7,8,10–12,14,16,17
R22	24R	Right	70 yrs	5 mos	2 mos	Noise/ sudden	29	4–6,9,12,14–18

[a]Across electrodes tested (in CL units).
[b]Bilaterally implanted.

duration was 500 ms with a 500-ms silent interval between stimuli. Stimulation was presented in monopolar mode relative to the extracochlear ground electrode MP1.

2. Procedure

Prior to behavioral testing, electrode impedance was measured to identify any electrode(s) with a short or open circuit so that malfunctioning electrodes would not be used in the experiment. Behavioral dynamic range was measured to determine stimulus levels for both ECAP and electrode pitch-ranking measures. First, *C* levels were measured across all viable electrodes in the array. Subjects were asked to indicate when the sound was loud but not uncomfortable. Repeated ascending trials were used, beginning with a step size of 5 current level units (CL) for the first pass, then 1-CL steps until the subject consistently identified the same *C* level on at least two of three trials. Levels were then balanced by sweeping across five electrodes at a time at *C* level with each group of five overlapping by one or two electrodes. The subject was instructed to identify any electrode that was perceived as unequal in loudness. *C* level for any electrode identified as noticeably softer or louder was raised or lowered, respectively, by 1 CL and the short sweep was repeated. This procedure was repeated until the subject reported that all electrodes were equal in loudness. If *C* level adjustments were made, they typically did not exceed more than 2–3 CL. A final sweep was performed across all electrodes in the array to verify equal loudness. This rather quick balancing procedure was used to obtain dynamic range levels that were relatively equal in loudness across all test electrodes. A meticulous loudness-balancing procedure was not employed because stimuli for the electrode pitch discrimination task were jittered in loudness at 70%, 80%, and 90% of these dynamic range measures to control for loudness cues.

T levels were then determined by ascending in 5-CL steps with approximately three stimulus presentations at each level until threshold was initially identified. A bracketing procedure was then used to find final threshold. Descending step size was 6–11 CL and ascending step size was 1 CL. *T* level was determined as the lowest level indicated by the subject on at least two of three trials. *T* levels were then balanced by sweeping across the electrode array at 50% of the dynamic range as described above for *C* levels. Subjects typically did not require adjustments during this phase. A final sweep was performed across all electrodes from apex to base at 100% of the dynamic range (*C* level) to verify pitch order of the electrodes. Sweeps were repeated as needed. All subjects reported appropriate ordering of pitch when these sweeps were performed. Test electrodes for each subject were chosen to include those with the highest and lowest *T* levels, as well as electrodes that were relatively evenly spaced across the array. Dynamic range was calculated for each test electrode to determine stimulus levels for ECAP and electrode pitch-ranking measures.

C. ECAP measures

1. Stimuli and procedure

ECAP measures were recorded using the Neural Response Telemetry (NRT, v. 3.0) software from Cochlear Corporation. Stimuli were delivered through the laboratory's SPrint body-worn speech processor interfaced with either a processor control interface (PCI) or portable programming system (PPS). Default parameters were used for rate (80 Hz), pulse width (25 μs/phase), gain (60 dB), masker/probe reference electrode (MP1), and recording reference electrode (MP2). Responses consisted of 50–200 averages.

NRT utilizes forward masking and a subtraction technique to reduce stimulus artifact from the ECAP response (Abbas *et al.*, 1999, 2004; Brown *et al.*, 1998, 2000; Cohen *et al.*, 2003; Dillier *et al.*, 2002). This method can also be used to assess interaction between neural populations that are stimulated by different electrodes (Abbas *et al.*, 2004; Cohen *et al.*, 2003; Hughes and Abbas, 2006). A companion paper illustrates this method in detail (Hughes and Abbas, 2006). Briefly, a probe pulse is fixed on one electrode, while a masker is systematically applied to each electrode in the array. With the subtraction method, maximum ECAP amplitude is typically produced when the masker and probe are on the same electrode. As the spatial separation between masker and probe increases, the subtracted ECAP response de-

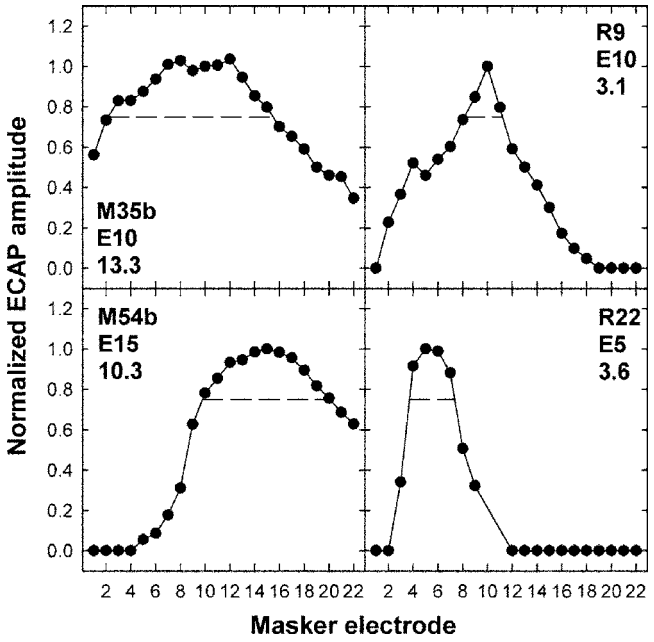


FIG. 1. Examples of ECAP masking functions that plot normalized amplitude as a function of masker electrode. Dashed lines represent width of the ECAP masking function at 75% normalized amplitude. Subject number, probe electrode number, and width (in number of electrodes) are denoted in each graph.

creases, indicating little or no interaction between masker and probe electrodes. The amplitude of the subtracted response therefore reflects the amount of spatial overlap (i.e., amount of interaction) between masker and probe.

For ECAP measures in this study, masker and probe levels were fixed at 80% of the behavioral dynamic range obtained with the 250-pps psychophysical stimulus for each electrode. These levels were the same for the electrode pitch-ranking portion of this study. For subjects M24 and M50, the 80% level yielded poor ECAP responses, so a 90% level was used.

2. Data analysis

ECAP amplitudes were calculated from the first negative peak (N1) to the following positive peak or plateau (P2) using a custom MATLAB computer program. If an N1 peak was not initially seen, recording parameters were adjusted until an N1 peak was obtained. Adjustments generally consisted of first changing recording delay or subsequently changing recording electrode. N1 peaks were visualized for all data reported here. Peak-to-peak amplitudes were plotted as a function of masker electrode and then normalized relative to the amplitude from the masker-equals-probe-electrode condition. The width of the ECAP amplitude function was measured in number of electrodes at 75% of the normalized ECAP amplitude. This level was chosen as it was the lowest percentage of normalized amplitude where the width could successfully be measured for all electrodes. Electrodes were not included in this study if a 75% width could not be obtained. The endpoints used to calculate width at 75% of the normalized amplitude were derived using the slope between adjacent electrodes' data points. Figure 1 shows sample ECAP masking patterns illustrating the variability across individuals and electrodes. In each graph, the horizontal dashed line represents the width of the function at 75% normalized amplitude (width in number of electrodes is denoted on each graph beneath probe electrode number). Typical response curves had largest amplitudes at or near the probe electrode, with decreasing amplitude as the distance between masker and probe electrodes increased. Examples of four different patterns are shown in Fig. 1. The top left panel shows a broad masking pattern with measurable masking on all electrodes (i.e., the function does not reach zero for any masker electrode). The bottom left panel shows a relatively broad masking pattern at 75% normalized amplitude; however, this function does reach zero at distal masker electrodes, indicating no masking effects for basal electrodes. The top right panel shows a narrow width at 75% normalized amplitude, but a significant amount of masking overall. The bottom right panel also shows a narrow width at 75%, but a very focused pattern emerges with no measurable masking on a total of 13 electrodes distal to the probe.

D. Electrode pitch ranking

1. Stimuli and procedure

A two-interval, two-alternative forced choice (2I2AFC) pitch-ranking task was used to test subjects' ability to discriminate between electrodes in the basis of pitch. The stimulus was a 250-pps, 500-ms train of 25-μs/phase biphasic pulses presented in monopolar mode (MP1), and was identical to the stimulus used to determine behavioral dynamic range. Stimuli and procedures are described in greater detail in a companion paper (Hughes and Abbas, 2006). Briefly, the first (reference) electrode stimulated was always the same within a trial. A 500-ms silent interval preceded stimulation of the second (test) electrode. The location of the test electrode was randomized and varied from one, two, four, or six electrode positions basal to and apical to the standard electrode, for a total of eight electrode pairs. The exception was for electrodes that were located less than six positions away from the basal or apical end or for short-circuit electrodes. For example, test electrodes for standard electrode 5 were electrodes 1, 2, 3, 4, 6, 7, 9, and 11. The subject was asked to indicate whether the second sound was higher or lower in pitch relative to the first. The procedure was similar to that described by Laneau and Wouters (2004). Presentation levels were randomized separately for each electrode at 70%, 80%, and 90% of the dynamic range to control for loudness cues. For the two subjects whose ECAP data were collected at 90% of the dynamic range, electrode pitch-ranking stimuli were presented at 80%, 90%, and 100% of the dynamic range. Each trial consisted of 72 presentations: one standard electrode at three levels multiplied by eight test electrodes at three levels each. Each trial was tested twice for every standard electrode. Presentation order of each trial was also randomized.

2. Data analysis

Results for a given electrode pair were pooled across the two trials and then quantified as a percent value for the number of times the subject responded that the second presenta-

tion was judged as higher in pitch relative to the first. Percent values were converted to z scores and plotted as a function of test electrode. Once a score reached 100% or 0%, subsequent scores on more basal or apical electrodes were eliminated to avoid falsely reducing the slope of the psychometric function. Simple linear regression analysis was then applied, and electrode pitch-ranking ability was quantified as the slope of the regression line.

III. RESULTS

A. Contour vs straight array

The primary goal of this study was to compare behavioral threshold (T level), ECAP channel interaction, and electrode pitch-ranking ability between individuals implanted with a traditional straight electrode array versus a perimodiolar array. It was expected that subjects implanted with the perimodiolar Contour device would have lower T levels, narrower ECAP masking functions (i.e., better spatial selectivity), and steeper slopes of the electrode pitch-ranking function when compared to subjects implanted with the traditional 24M straight array. Two-way analysis of variance (ANOVA) was performed for each measure (T level, width of the ECAP function, and slope of the z score function) to assess the effect of device type and electrode place. Electrode place was characterized as basal (electrodes 1–8), middle (electrodes 9–15), or apical (electrodes 16–22). Because each subject had a different number of electrodes tested within each electrode place grouping, measures from all electrodes within a place grouping were averaged together for each subject for statistical analyses reported in the following three sections.

1. T level

Figure 2(a) shows a comparison of behavioral T levels across electrodes for the two device types. Filled circles represent mean data (±1 standard error, or SE) for each probe electrode in subjects with the 24M straight array. Open circles represent mean data (±1 SE) for each electrode in subjects with the 24R Contour array. Two-way ANOVA results showed no significant effect of device type or electrode place on T level ($p>0.1$). There was no significant interaction effect between device type and electrode place ($p>0.1$). Mean T level for subjects with the straight array was 162.3 CL (SE=1.6), compared to 158.9 CL (SE=2.2) for Contour users. This outcome is consistent with previously presented data from a larger set of subjects from this laboratory (Brown *et al.*, 2002; Hughes *et al.*, 2001). However, one would expect to see a significant difference in T level between the two groups based on the theory behind differences in implant design and previous research from other laboratories showing lower T levels for perimodiolar devices (i.e., Cohen *et al.*, 2001; Kawano *et al.*, 1998; Kreft *et al.*, 2004; Parkinson *et al.*, 2002; Saunders *et al.*, 2002; Young and Grohne, 2001).

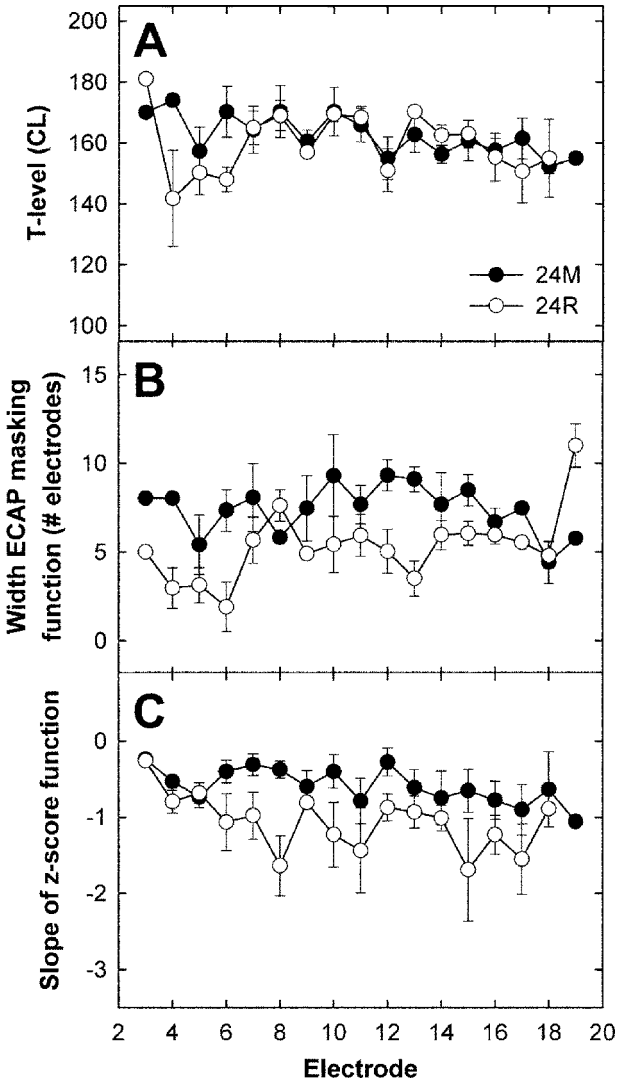


FIG. 2. (A) Group data showing mean behavioral threshold (T level) across electrodes for the two device types. (B) Group data showing mean width of the ECAP masking function across electrodes for the two device types. (C) Group data showing slope of the electrode pitch-ranking function across electrodes for the two device types. Filled circles represent average data from subjects with the 24M straight array. Open circles represent average data from subjects with the 24R Contour array. Error bars represent 1 standard error around the mean.

2. ECAP masking patterns

Figure 2(b) shows average width of the ECAP masking functions at 75% normalized amplitude for the two device types. Two-way ANOVA results showed a significant effect of device type on width of the ECAP masking function. Specifically, the width of the ECAP masking function was significantly narrower overall for subjects implanted with the Contour array compared to subjects implanted with the straight array ($p=0.007$). There was no significant effect of electrode place ($p>0.1$) or interaction effects between device and electrode place ($p>0.1$). *Post hoc* pairwise comparisons using the Holm-Sidak method showed significantly narrower ECAP masking functions for the Contour than for the straight array for basal ($p=0.025$) and middle ($p=0.045$) electrode regions, but not for the apical region ($p=0.53$). Mean width across all electrodes for all subjects with

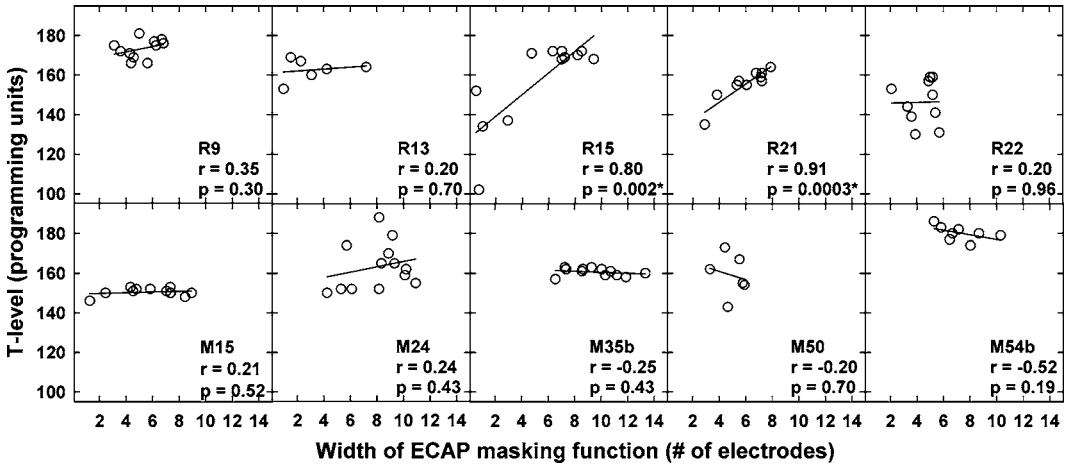


FIG. 3. Individual results comparing width of the ECAP masking function vs behavioral threshold (T level). Each panel represents data from a different subject. Solid lines represent results of linear regression analyses. Subject number is indicated on each graph, followed by correlation coefficient and p value. Asterisks indicate statistical significance ($p<0.05$).

the straight array was 7.4 (SE=0.35) electrodes, compared to 5.0 (SE=0.31) electrodes for Contour users. These results suggest that there is less physiologic channel interaction and better spatial selectivity for subjects implanted with the Contour array compared to the straight array.

3. Electrode pitch ranking

Figure 2(c) shows the average slopes of the psychophysical electrode pitch-ranking functions for the two device types. Two-way ANOVA results showed a significant effect of device type on slope of the electrode pitch-ranking function. Specifically, Contour subjects had steeper slopes than straight-array subjects ($p=0.039$). There was no significant effect of electrode place and no interaction effects between electrode place and device type ($p>0.1$). Mean slope for subjects with the straight array was −0.61 (SE=0.06), compared to −1.10 (SE=0.09) for Contour users. These results suggest that on average, the subjects in this study who were implanted with the Contour array performed better on an electrode pitch-ranking task when compared to subjects implanted with the straight array. It is worthwhile to note that mean duration of implant use at the time of testing was 3.9 years in the straight-array group vs 1.4 years in the Contour group. Given that the better-performing Contour group had significantly less experience with their implants than the straight-array group, it is unlikely that duration of device use was a critical factor influencing performance on the electrode pitch-ranking task. Additionally, while the average age of the Contour subjects was relatively lower (mean=55.6 years) compared to the straight-array group (mean=62.6 years), the difference in age between the two groups was not statistically significant ($p>0.05$).

B. *T*-level vs ECAP masking

1. Individual results

Figure 3 shows the relation between width of the ECAP masking function and T level for each of the ten subjects. It was hypothesized that sharper ECAP masking patterns (e.g., better spatial selectivity represented as narrow width measures) would be recorded for electrodes with lower T levels, and broader ECAP masking patterns (larger widths) would result from electrodes with higher T levels. Thus, the regression line in these scatter plots should have a positive slope. A strong positive correlation was found for two Contour subjects, R15 ($r=0.80$, $p=0.002$) and R21 ($r=0.91$, $p=0.0003$). These results were statistically significant and were consistent with the hypothesis. There was no statistically significant relation between T level and width of the ECAP masking function for the remaining eight subjects. Correlation coefficients and p values are reported in each panel of Fig. 3 beneath the subject number. It is of interest to note that of the two Contour subjects with a significant correlation between T level and width of the ECAP masking function, one subject (R15) performed well on speech perception tests (e.g., 96% on HINT sentences) and the other subject (R21) did not perform as well (e.g., 49% on HINT sentences). Speech perception scores for each subject are shown in Table II. (HINT sentence scores were an average of four lists; CNC word and phoneme scores were an average of two lists. All speech perception tests were presented in quiet using the subject's own processor with most-used settings.) Further comparisons between speech perception performance and ECAP channel

TABLE II. Speech perception scores (in percent correct) for subjects participating in this study.

Subject number	HINT sentences	CNC words	CNC phonemes
M15	95%	63%	82%
M24	81%	61%	80%
M35b[a]	95%	48%	72%
M50	53%	42%	63%
M54b[a]	95%	40%	66%
R9	85%	22%	48%
R13	85%	31%	58%
R15	96%	57%	77%
R21	49%	3%	11%
R22	82%	34%	53%

[a]Bilaterally implanted.

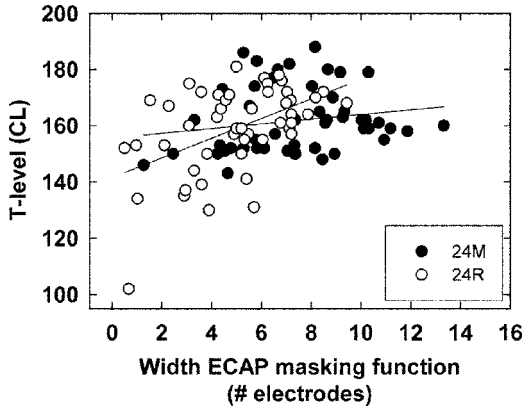


FIG. 4. Group data showing the relation between width of the ECAP masking function and behavioral threshold (T level). Filled circles represent individual electrode data from subjects with the 24M straight array. Open circles represent data from subjects with the 24R Contour array. Linear regression lines are plotted separately for each device type.

interaction or electrode pitch-ranking ability are described in a companion paper (Hughes and Abbas, 2006), which showed no significant relation between speech perception performance and average width of the ECAP masking function or electrode pitch ranking for these subjects.

2. Group results

Figure 4 shows the relation between T level and width of the ECAP function for all electrodes in all subjects. Again, one would expect low T levels for narrow ECAP masking functions and high T levels for broader ECAP masking functions (i.e., positive regression line slope). When all data from both subject groups were pooled, a statistically significant positive correlation was found ($r=0.34$, $p<0.001$), which is consistent with the hypothesis. Thus, electrodes with better spatial selectivity (narrow ECAP width measures) tended to have lower T levels and electrodes with poorer spatial selectivity (large width measures) had higher T levels. When data were analyzed as two separate groups according to device type, both groups showed trends in the same direction; however, only data from the Contour subjects (open circles) were significant ($r=0.49$, $p<0.001$, Contour array; $r=0.18$, $p=0.21$, straight array). These data suggest better spatial selectivity exists for electrodes with lower T levels in subjects implanted with the Contour array.

C. *T* level vs electrode pitch ranking

1. Individual results

Figure 5 shows the relation between electrode pitch-ranking ability (slope of the psychophysical function) and T level for each of the ten subjects. Data are presented in a format similar to Fig. 3. It was hypothesized that better electrode pitch-ranking ability (i.e., larger negative slope values) would occur for electrodes with lower T level, and poorer electrode pitch-ranking ability (shallower slopes) would occur for electrodes with higher T level. Thus, regression lines should have a positive slope. Results in Fig. 5 show that there was not a single subject who demonstrated a significant trend in the predicted direction. Two subjects (R13 and M50) showed a strong correlation in the opposite direction ($r=-0.89$ and $r=-0.76$); however, only the data from subject R13 were statistically significant ($p=0.02$). Correlation coefficients and p values for Fig. 5 data are listed in each panel beneath the subject number.

Results from Fig. 5 indicate that low T levels are not correlated with better electrode pitch-ranking ability, as proposed in this study. In fact, a statistically significant trend in the opposite direction was found for one subject. Results from roughly one-third of the subjects (R15, R22, M50) showed a trend similar to that subject; however, those results were not statistically significant ($p=0.34$, $p=0.21$, $p=0.08$, respectively).

2. Group results

Figure 6 shows the relation between T level and slope of the electrode pitch-ranking function for all subjects participating in this study. Again, one would expect to see a positive slope of the regression line, indicating low T levels are associated with better electrode pitch-ranking ability (steeper functions) and high T levels associated with poorer electrode

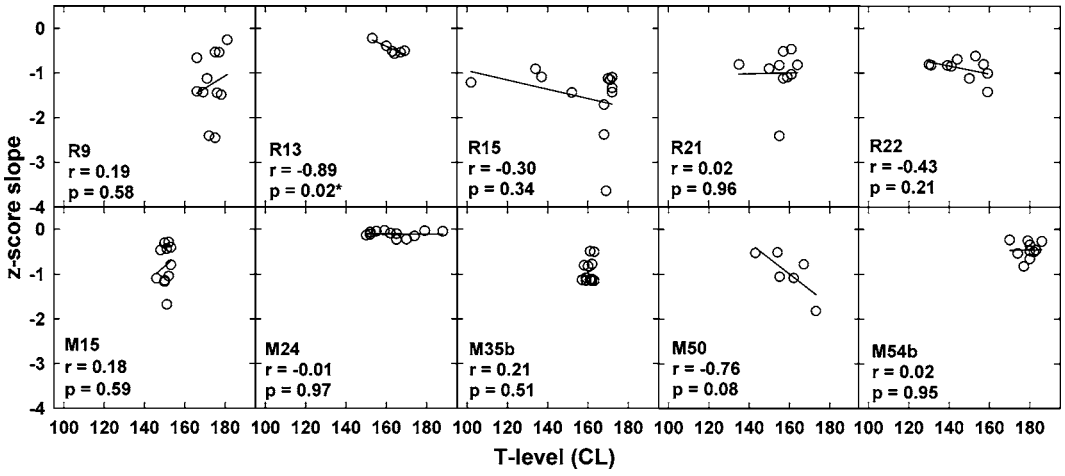


FIG. 5. Individual results comparing behavioral threshold (T level) vs slope of the electrode pitch-ranking function. Each panel represents data from a different subject. Solid lines represent results of linear regression analyses. Subject number is indicated on each graph followed by correlation coefficient and p value. Asterisk indicates statistical significance ($p<0.05$).

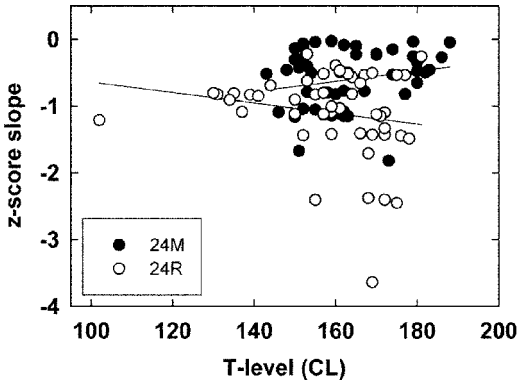


FIG. 6. Group data comparing slope of electrode pitch-ranking function to behavioral threshold (*T* level). Filled circles represent individual electrode data from subjects with the 24M straight array. Open circles represent individual electrode data from subjects with the 24R Contour array. Linear regression lines are plotted separately for each device type.

pitch-ranking ability (shallow functions). When all data from both subject groups were pooled, there was no significant correlation between these two measures ($r=0.01$, $p=0.93$). Similarly, there was no correlation between the two measures when individual data were analyzed as a function of device type ($r=0.20$, $p=0.15$, straight array; $r=-0.19$, $p=0.20$, Contour array). Recall that individual data from Fig. 5 showed a significant negative correlation between these two measures for one subject and similar trends (although not significant) for one third of the remaining subjects, indicating a trend opposite of that hypothesized. This trend did not hold for group data, and suggests that there is no relation between *T* level and electrode pitch-ranking ability, regardless of device type. These results were consistent with those reported by Pfingst *et al.* (1999).

IV. DISCUSSION

A. Contour vs straight array

The primary goal of this study was to compare *T* level, width of the ECAP masking function, and electrode pitch-ranking ability for subjects with a traditional straight electrode array (Nucleus 24M) vs an array designed for perimodiolar positioning [Nucleus 24R(CS) Contour]. It was hypothesized that subjects with the Contour array would have lower *T* levels, less ECAP channel interaction, and better electrode pitch-ranking ability compared to subjects with the straight array. Results from this study showed that there was no difference in *T* level between device types; however, Contour subjects had significantly narrower ECAP masking widths and better electrode pitch-ranking ability than subjects implanted with the straight array. Unfortunately, these significant differences did not translate into significantly better speech perception performance for the Contour group (Hughes and Abbas, 2006).

Although Fig. 2(a) shows lower *T* levels for Contours on three basal electrodes, this difference was not statistically significant due to the small number of observations for some of the electrodes. Overall, there was no significant difference in *T* level between the two device types across the array. This finding is in contrast to other studies reporting lower *T* levels in the Contour vs the 24M straight array (e.g., Parkinson *et al.*, 2002; Saunders *et al.*, 2002). There are a number of possibilities to which we may attribute this lack of significant difference. First, there may not actually be a difference in modiolus proximity between the two array types. Radiological data were not available to confirm or refute this hypothesis. However, given the precurved design of the Contour array, it is unlikely that the Contour arrays consistently occupied a lateral wall position as straight arrays tend to. Several studies have used various imaging techniques to determine electrode position within the cochlea for different array designs. These studies demonstrated that straight arrays typically occupy a lateral wall position and precurved arrays occupy either an intermediate or perimodiolar position (Cohen *et al.*, 2001, 2003; Roland *et al.*, 2000; Saunders *et al.*, 2002). It may be that electrode position for the Contour subjects in this study was more intermediate than perimodiolar.

A second possible reason for the lack of difference in *T* level between the two device types may be due to a combination of differences in electrode design and electrode placement. Electrodes in the 24M straight array consist of full platinum bands that encompass the entire circumference of the silastic carrier. In contrast, the Contour array consists of half-band electrodes located on the medial portion of the precurved silastic carrier. Thus, orientation and spread of current fields will be different for the two electrode designs. Broader stimulation patterns are known to significantly reduce *T* level, as demonstrated by research comparing bipolar to monopolar stimulation, for example (Clark *et al.*, 1988; Miller *et al.*, 1999; Pfingst *et al.*, 1979, 1995, 1997; Pfingst, Miller *et al.*, 1995; Pfingst and Sutton, 1984; Shannon, 1983; Smith *et al.*, 1994; Smith and Finley, 1997). If current spread is broader with the full-band electrode relative to the half-band electrode, then one would expect a lower threshold with the full-band design for an equivalent electrode placement. The presumed closer placement of the half-band electrode array to the modiolus and neural elements may compensate for that difference to some extent. Thus, the combination of lateral wall location and broad current fields produced by the straight array could yield similar behavioral thresholds as the combined effects of less-than-optimal (intermediate) electrode positioning and smaller current fields produced by the Contour array. It is not clear, however, why electrode position for subjects in the present study would differ significantly from other subjects with the same two device types compared in other studies (e.g., Parkinson *et al.*, 2002; Saunders *et al.*, 2002).

Although *T* levels were not significantly different between the two device types, Contour subjects demonstrated significantly better electrode pitch-ranking ability and narrower ECAP masking functions than subjects with the straight array. If modiolus proximity is responsible for less ECAP channel interaction and better electrode pitch-ranking ability, then this result suggests that, in contrast to Kawano *et al.* (1998) and Cohen *et al.* (2001), *T* level may not in fact be a very good predictor of modiolus proximity. It is possible, however, that there were differences in *T* level between the

two device types, but that the differences were not large enough to be statistically significant for this group of subjects.

Another explanation for this interesting finding may be a combination of electrode design and stimulus levels used for each measure. Clearly, *T* level is a threshold measure, whereas both the ECAP and electrode pitch-ranking measures were suprathreshold (80%–90% of the dynamic range). It is possible that the different electrode designs (full vs half band) yield different current field patterns with effects that are more pronounced at higher levels. Specifically, the half-band electrode design in the Contour may yield more focused current fields at higher levels than full bands, resulting in less physiologic interaction and better electrode pitch-ranking ability. It should be noted that there was no significant difference in *C* levels between the two array types ($p = 0.54$). This finding is in contrast to a number of studies reporting lower *C* levels for perimodiolar electrodes (Kawano *et al.*, 1998; Kreft *et al.*, 2004; Parkinson *et al.*, 2002; Saunders *et al.*, 2002; Young and Grohne, 2001). Thus, better electrode pitch-ranking ability and less ECAP masking in the Contour subjects cannot be attributed to differences in the overall amount of current used. This is consistent with reports from McKay *et al.* (1999) and Pfingst *et al.* (1999) that showed little effect of stimulus level on electrode discrimination ability. It remains unclear exactly why differences emerged between device types for electrode pitch-ranking ability and ECAP channel interaction, but not for *T* level.

B. *T* level vs ECAP masking

A second goal of this study was to compare *T* level to width of the ECAP masking function. It was hypothesized that electrodes with lower *T* levels would exhibit less ECAP channel interaction within and across subjects. Results showed a significant positive correlation between *T* level and ECAP masking width for two Contour subjects, as well as for all Contour group data. As Fig. 4 shows, the largest variance in *T* level occurred for Contour subjects. Thus, it appears that the relation between *T* levels and ECAP channel interaction can primarily be attributed to differences across subjects, as only two of the ten subjects exhibited this trend in the individual data. Although lower *T* levels can result from either better neural survival or electrode proximity to neural elements, it would be expected that lower *T* levels would be correlated with more discrete excitation patterns within the cochlea because either the electric fields are steeper and narrower (for proximal electrode placements) or because there is a denser neural population.

C. *T* level vs electrode pitch ranking

A third goal of this study was to compare *T* level with slope of the electrode pitch-ranking function. It was hypothesized that electrodes with lower *T* levels would yield better electrode pitch-ranking ability within and across subjects. Results showed a significant negative correlation for one Contour subject, which was opposite the hypothesis. No significant results were found for group data across all electrodes. Overall, there is little evidence to suggest that lower *T* levels are indicative of better electrode pitch-ranking ability. These results are consistent with those reported by Pfingst *et al.* (1999). This outcome may be because one measure is threshold and the other is suprathreshold; that is, *T* level is not predictive of current spread patterns at higher stimulus levels. However, Pfingst *et al.* (1999) and McKay *et al.* (1999) showed only small effects of stimulus level for electrode discrimination tasks. They proposed that electrode discrimination might be attributed more to shifts in the edges of excitation patterns rather than the overall degree of overlap between populations stimulated by different electrodes. However, we might still expect that even the edges of excitation patterns would vary more between electrodes if stimulating current fields were more selective. Another possible reason for the lack of correlation is that *T* level is a simple behavioral detection measure, whereas the electrode pitch-ranking task also requires processing and decision making, both of which are higher-level tasks. This latter explanation may also account for results reported in a companion paper (Hughes and Abbas, 2006), which showed no significant correlation between width of the ECAP masking function and slope of the electrode pitch-ranking function for these subjects.

V. CONCLUSIONS

This study showed that Contour subjects had significantly better electrode pitch-ranking ability and less ECAP channel interaction compared to straight-array subjects, despite similar behavioral thresholds in the two device types. These findings might be attributed to differences in electrode contact design rather than modiolus proximity (or neural survival), or differences in stimulus level (threshold vs suprathreshold) used for each measure.

ACKNOWLEDGMENTS

This study is based on doctoral dissertation work at the University of Iowa by the first author. The second author served as adviser and chair of the dissertation committee. This study was supported by funds from the NIH/NIDCD (Grant P50 DC00242), NIH/NCRR (Grant RR00059), and an anonymous donor. The authors would like to thank George Woodworth for assistance with statistical analyses, and Arik Wald for programming support. Additional thanks go to the ten subjects who participated in this study, and to two anonymous reviewers for helpful comments.

Abbas, P. J., Brown, C. J., Shallop, J. K., Firszt, J. B., Hughes, M. L., Hong, S. H., and Staller, S. J. (**1999**). "Summary of results using the Nucleus CI24M implant to record the electrically evoked compound action potential," Ear Hear. **20**, 45–59.

Abbas, P. J., Hughes, M. L., Brown, C. J., Miller, C. A., and South, H. (**2004**). "Channel interaction in cochlear implant users evaluated using the electrically evoked compound action potential," Audiol. Neuro-Otol. **9**, 203–213.

Briaire, J. J., and Frijns, J. H. M. (**2000**). "Field patterns in a 3D tapered spiral model of the electrically stimulated cochlea," Hear. Res. **148**, 18–30.

Brown, C. J., Abbas, P. J., and Gantz, B. J. (**1998**). "Preliminary experience with Neural Response Telemetry in the Nucleus CI24M cochlear implant,"

Am. J. Otol. **19**, 320–327.
Brown, C. J., Hughes, M. L., Abbas, P. J., South, H., and Schmidt-Clay, K. (**2002**). "Comparisons between NRT thresholds and 900-Hz MAP T- and C-levels in Nucleus cochlear implant users," 7th International Cochlear Implant Conference, Manchester, United Kingdom, September 2002, p. 43.
Brown, C. J., Hughes, M. L., Luk, B., Abbas, P. J., Wolaver, A. A., and Gervais, J. P. (**2000**). "The relationship between EAP and EABR thresholds and levels used to program the Nucleus CI24M speech processor: Data from adults," Ear Hear. **21**, 151–163.
Clark, G. M., Shepherd, R. K., Franz, B. K.-H., Dowell, R. C., Tong, Y. C., Blamey, P. J., Webb, R. L., Pyman, B. C., McNaughtan, J., Bloom, D. M., Kakulas, B. A., and Siejka, S. (**1988**). "The histopathology of the human temporal bone and auditory central nervous system following cochlear implantation in a patient," Acta Oto-Laryngol., Suppl. **448**, 1–65.
Cohen, L. T., Richardson, L. M., Saunders, E., and Cowan, R. S. C. (**2003**). "Spatial spread of neural excitation in cochlear implant recipients: Comparison of improved ECAP method and psychophysical forward masking," Hear. Res. **179**, 72–87.
Cohen, L. T., Saunders, E., and Clark, G. M. (**2001**). "Psychophysics of a prototype perimodiolar cochlear implant electrode array," Hear. Res. **155**, 63–81.
Dillier, N., Lai, W. K., Almqvist, B., Frohne, C., Muller-Deile, J., Stecker, M., and von Wallenberg, E. (**2002**). "Measurement of the electrically evoked compound action potential via a neural resonse telemetry system," Ann. Otol. Rhinol. Laryngol. **111**, 407–414.
Finley, C. C., Wilson, B. S., and White, M. W. (**1990**). "Models of neural responsiveness to electrical stimulation," in *Cochlear Implants Models of the Electrically Stimulated Ear*, edited by J. M. Miller, and F. A. Spelman (Springer, New York), pp. 55–96.
Frijns, J. H. M., de Snoo, S. L., and Schoonhoven, R. (**1995**). "Potential distribution and neural excitation patterns in a rotationally symmetric model of the electrically stimulated cochlea," Hear. Res. **87**, 170–186.
Frijns, J. H. M., de Snoo, S. L., and ten Kate, J. H. (**1996**). "Spatial selectivity in a rotationally symmetric model of the electrically stimulated cochlea," Hear. Res. **95**, 33–48.
Hughes, M. L., and Abbas, P. J. (**2006**). "The relation between electrophysiologic channel interaction and electrode pitch ranking in cochlear implant recipients," J. Acoust. Soc. Am. **119**, 1527–1537.
Hughes, M. L., Abbas, P. J., Brown, C. J., Seyle, K., and South, H. (**2001**). "Assessing spatial spread in straight and modiolar hugging electrode arrays," 8th Symposium on Cochlear Implants in Children, February 2001, Los Angeles, CA, p. 55.
Kawano, A., Seldon, H. L., Clark, G. M., Ramsden, R. T., and Raine, C. H. (**1998**). "Intracochlear factors contributing to psychophysical percepts following cochlear implantation," Acta Otolaryngol. Stockh. **118**, 313–326.
Kral, A., Hartmann, R., Mortazavi, D., and Klinke, R. (**1998**). "Spatial resolution of cochlear implants: The electrical field and excitation of auditory afferents," Hear. Res. **121**, 11–28.
Kreft, H. A., Donaldson, G. S., and Nelson, D. A. (**2004**). "Effects of pulse rate on threshold and dynamic range in Clarion cochlear implant users," J. Acoust. Soc. Am. **115**, 1885–1888.
Laneau, J., and Wouters, J. (**2004**). "Multichannel place pitch sensitivity in cochlear implant recipients," J. Assoc. Res. Otolaryngol. **5**, 285–294.
Liang, D. H., Lusted, H. S., and White, R. L. (**1999**). "The nerve-electrode interface of the cochlear implant: Current spread," IEEE Trans. Biomed. Eng. **46**, 35–43.
McKay C. M., O'Brien A., James C. J., (**1999**). "Effect of current level on electrode discrimination in electrical stimulation," Hear. Res. **136**, 159–164.
Miller, A. L., Smith, D. W., and Pfingst, B. E. (**1999**). "Across-species comparisons of psychophysical detection thresholds for electrical stimulation of the cochlea. II. Strength-duration functions for single, biphasic pulses," Hear. Res. **135**, 47–55.
Parkinson, A. J., Arcaroli, J., Staller, S. J., Arndt, P. L., Cosgriff, A., and Ebinger, K. (**2002**). "The Nucleus 24 Contour cochlear implant system: Adult clinical trial results," Ear Hear. **23**, 41S–48S.
Pfingst, B. E., Donaldson, J. A., Miller, J. M., and Spelman, F. A. (**1979**). "Psychophysical evaluation of cochlear prostheses in a monkey model," Ann. Otol. Rhinol. Laryngol. **88**, 613–625.
Pfingst, B. E., Holloway, L. A., Zwolan, T. A., and Collins, L. M. (**1999**). "Effects of stimulus level on electrode-place discrimination in human subjects with cochlear implants," Hear. Res. **134**, 105–115.
Pfingst, B. E., Miller, A. L., Morris, D. J., Zwolan, T. A., Spelman, F. A., and Clopton, B. M. (**1995**). "Effects of electrical current configuration on stimulus detection," Ann. Otol. Rhinol. Laryngol. Suppl. **166**, 127–131.
Pfingst, B. E., Morris, D. J., and Miller, A. L. (**1995**). "Effects of electrode configuration on threshold functions for electrical stimulation of the cochlea," Hear. Res. **85**, 76–84.
Pfingst, B. E., and Sutton, D. (**1984**). "Relation of psychophysical thresholds for electrical stimuli to auditory nerve survival: Summary of results for 18 scala tympani implants," Abstracts of the 7th Midwinter Research Meeting of the Association for Research in Otolaryngology, p. 10.
Pfingst, B. E., Zwolan, T. A., and Holloway, L. A. (**1997**). "Effects of stimulus configuration on psychophysical operating levels and on speech recognition with cochlear implants," Hear. Res. **112**, 247–260.
Ranck, J. B. (**1975**). "Which elements are excited in electrical stimulation of mammalian central nervous system: A review," Brain Res. **98**, 417–440.
Rattay, F., Leao, R. N., and Felix, H. (**2001**). "A model of the electrically excited human cochlear neuron. II. Influence of the three-dimensional cochlear structure on neural excitability," Hear. Res. **153**, 64–79.
Roland, J. T., Fishman, A. J., Alexiades, G., and Cohen, N. L. (**2000**). "Electrode to modiolus proximity: A fluoroscopic and histologic analysis," Am. J. Otol. **21**, 218–225.
Saunders, E., Cohen, L., Aschendorff, A., Shapiro, W., Knight, M., Stecker, M., Richter, B., Waltzman, S., Tykocinski, M., Roland, T., Laszig, R., and Cowan, R. (**2002**). "Threshold, comfortable level, and impedance changes as a function of electrode-modiolar distance," Ear Hear. **23**, 28S–40S.
Shannon, R. V. (**1983**). "Multichannel electrical stimulation of the auditory nerve in man. I. Basic psychophysics," Hear. Res. **11**, 157–189.
Shepherd, R. K., Hatsushika, S., and Clark, G. M. (**1993**). "Electrical stimulation of the auditory nerve: The effect of electrode position on neural excitation," Hear. Res. **66**, 108–120.
Smith, D. W., and Finley, C. C. (**1997**). "Effects of electrode configuration on psychophysical strength-duration functions for single biphasic electrical stimuli in cats," J. Acoust. Soc. Am. **102**, 2228–2237.
Smith, D. W., Finley, C. C., van den Honert, C., Olszyk, V. B., and Konrad, K. E. M. (**1994**). "Behavioral and electrophysiological responses to electrical stimulation in the cat. I. Absolute thresholds," Hear. Res. **81**, 1–10.
Young, N. M., and Grohne, K. M. (**2001**). "Comparison of pediatric Clarion recipients with and without the electrode positioner," Otol. Neurotol. **22**, 195–199.

Auditory evoked potentials in females with high and low acceptance of background noise when listening to speech

Joanna W. Tampas and Ashley W. Harkrider
Department of Audiology and Speech Pathology, University of Tennessee, 457 South Stadium Hall, Knoxville, Tennessee 37996

(Received 14 June 2005; revised 12 December 2005; accepted 27 December 2005)

Acceptable noise level (ANL) is a measure of a listener's acceptance of background noise when listening to speech. A consistent finding in research on ANL is large intersubject variability in the acceptance of background noise. This variability is not related to age, gender, hearing sensitivity, type of background noise, speech perception in noise performance, cochlear responses, or efferent activity of the medial olivocochlear pathway. In the present study, auditory evoked potentials were examined in 21 young females with normal hearing with low and high acceptance of background noise to determine whether differences in judgments of background noise are related to differences measured in aggregate physiological responses from the auditory nervous system. Group differences in the auditory brainstem response, auditory middle latency response, and cortical, auditory late latency response indicate that differences in more central regions of the nervous system account for, at least in part, the variability in listeners' willingness to accept background noise when listening to speech. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2167147]

PACS number(s): 43.64.Qh, 43.64.Ri, 43.66.Ba [WPS] Pages: 1548–1561

I. INTRODUCTION

Nabelek *et al.* (1991) developed a procedure to quantify the maximum amount of background noise that listeners are willing to accept while listening to speech. In this procedure, listeners adjust the background noise level (BNL) to the highest intensity they deem acceptable while listening to and following the words of a recorded story at their most comfortable listening level (MCL). This measure, which has become known as acceptable noise level (ANL), evaluates an individual's willingness to accept noise while listening to speech (Nabelek *et al.* 2004). ANL is computed by subtracting the maximum acceptable background noise level an individual is willing to accept from the speech presentation level (e.g., ANL=MCL-BNL). Therefore, individuals with low ANLs (e.g., 6 dB) are willing to accept larger amounts of background noise than those with high ANLs (e.g., 16 dB).

Listeners' reactions to background noise are not uniform. Recent investigations of background noise acceptance have attempted to identify the variables contributing to the wide range of differences exhibited among individuals in their acceptance of background noise. Although ANLs have been found to be highly reliable within and between test sessions in individuals (Nabelek *et al.*, 2004), studies report large between-subject differences in the acceptance of background noise among homogeneous populations. These investigations have failed to explain the large range of ANLs on the basis of age, hearing sensitivity (Nabelek *et al.*, 1991), amplification with hearing aids (Nabelek *et al.*, 2004), gender (Rogers *et al.*, 2003), type of background noise (Nabelek *et al.*, 1991; Crowley and Nabelek, 1996), cochlear responses (Harkrider and Smith, 2005; Tampas *et al.*, 2004), efferent activity of the medial olivocochlear bundle pathway, middle ear characteristics (Harkrider and Smith, 2005), and speech perception in noise performance (Nabelek *et al.*, 2004). Moreover, across ANL studies, young and elderly individuals with both hearing impairment and normal hearing sensitivity display equivalent means and ranges for ANLs (for review, see Rogers *et al.*, 2003), indicating that acceptance of background noise may be an inherent characteristic of the individual that does not change with age or the development of hearing loss.

Nabelek *et al.* (2004) report that, on average, full-time hearing-aid users accept more background noise (i.e., have lower ANLs) than part-time hearing-aid users. However, the range of ANLs within the groups is relatively large and an understanding of the origin of the variability is needed to develop treatment strategies to increase acceptance of background noise, consequently increasing hearing-aid use. Thus, identification of the variables contributing to the large range of ANLs in listeners with normal hearing would have important implications for impaired hearing populations, possible treatment strategies, and would result in an overall improved understanding of the processes contributing to this behavioral phenomenon. Previous research (Harkrider and Smith, 2005; Tampas *et al.*, 2004) suggests that differences in background noise acceptance may be related to individual variations in the afferent and/or efferent function of the central auditory nervous system, as measured with auditory evoked potentials (AEPs), but this possibility has not been fully explored.

The human brain provides both inhibitory and excitatory input to maintain performance in various types of tasks by means of waves of neuronal hyperpolarization and depolarization. At synaptic junctions, these actions trigger the release of neurotransmitters to exert one of two effects on postsynaptic membranes: excitation, which decreases membrane thresholds and increases the probability of neuronal firing or inhibition, which raises membrane thresholds and decreases

 0001-4966/2006/119(3)/1548/14/$22.50

the probability of neuronal firing. These processes result in a progressive and sequential order of informational processing along the auditory afferent pathways, from the auditory nerve to the cerebral cortex. These ascending pathways are reciprocal, meaning that the auditory system is rich in efferent, descending projections. Therefore, the function of a given auditory nucleus can be influenced by feedback from higher auditory nuclei, ultimately the auditory cortex. Various AEPs allow gross measurement of the electrical activity from the central auditory system detected through electrodes placed on the scalp. These measures can be employed to describe shifts in the balance between neural excitation and inhibition (e.g., Ashmore, 1991; Chao and Knight, 1997; Dustman *et al.*, 1996; Knight *et al.*, 1999; McDowd and Filion, 1992; Lei-Zhang *et al.*, 2001).

Research has demonstrated that efferent feedback plays an instrumental role in physiologic/psychoacoustical phenomena. For example, the medial olivocochlear (MOC) system has been implicated in protecting the ear from acoustic injury: stimulation of the MOC bundle (MOCB) reduces temporary threshold shifts (TTSs) (Rajan, 1991), and severing the MOCB increases permanent threshold shifts (PTSs) (Kujawa and Liberman, 1997). In a recent study (Harkrider and Smith, 2005) of 30 young adults with normal hearing, ANLs were found to be unrelated to middle ear impedance measures, acoustic reflex thresholds evoked by tonal or broad band noise stimuli, or contralateral suppression of click-evoked otoacoustic emissions (CEOAEs). These authors concluded that variability in ANLs was not due to middle-ear characteristics or activity levels of the acoustic reflex or MOC system. However, monotic (typically recorded) ANL was correlated with dichotic ANL (story in one ear, background noise in opposite ear) implying ANL must be mediated, in part, beyond the level of the superior olivary complex where binaural processing first occurs in the auditory system. Furthermore, Freyaldenhoven *et al.* (2005) investigated the effects of stimulant medications on background noise acceptance in young adults with attention deficit/hyperactivity disorder (ADHD/ADD) and normal hearing sensitivity. The results revealed decreased ANLs with the introduction of medication, indicating that the individuals were more willing to accept background noise when medicated. The authors concluded that the reduction of ANL is most likely not due to changes in peripheral auditory phenomenon but a result of changes in central processing (i.e., suppression of cortical activity, enhanced inhibitory processes, etc.) when medicated.

In a preliminary study presented in abstract form, Tampas *et al.* (2004) obtained physiological responses from two groups of females [one with low ANLs ($n=7$) and one with high ANLs ($n=6$)], including otoacoustic emissions (OAEs), auditory brainstem responses (ABRs), and middle latency responses (MLRs). OAEs are low-level, narrowband acoustic signals found in the external auditory canal (EAC) that are a by-product of nonlinear mechanical activity within the cochlea (i.e., cochlear amplifiers) (Davis, 1983). The ABR is one type of AEP representing synchronized activity of the auditory nerve and brainstem structures (Moller, 1994) and occurring approximately 1–7 ms following the onset of a transient stimulus. It consists of a series of five to seven peaks described relative to their absolute latencies. These peaks are designated by roman numerals (I–VII), with the most prominent being waves I, III, and V (Hall, 1992). The auditory MLR is composed of a series of negative and positive peaks (Na, Pa, and Nb) that occur within approximately 10–75 ms of stimulus onset and immediately following the ABR (Jerger *et al.*, 1988; Picton *et al.*, 1974). Researchers have identified the underlying neuronal generators of the MLR to include subcortical (inferior colliculus, thalamus, thalamocortical pathway) and cortical (temporal lobe) structures (e.g., Hashimoto, 1982; Jacobson *et al.*, 1990). The data obtained by Tampas *et al.* (2004) suggested that more central regions of the auditory system might account for variability in listeners' willingness to accept background noise. Specifically, no differences emerged between the females with low versus high ANLs in OAEs, AEP latencies, or amplitudes of early waves I and III of the ABR response. However, group differences did emerge in the amplitudes of ABR wave V and MLR waves Na and Pa; the low ANL group, overall, tended to exhibit smaller amplitudes than the high ANL group. The results indicated that physiological differences in the auditory system were a potential variable contributing to the intersubject variability in background noise acceptance and that further investigation was needed to follow up on these group differences.

The main purpose of the current experiment was to enhance understanding of the regions in the auditory system that may be related to an individual's background noise acceptance when listening to speech. More specifically, to determine if individual differences in physiological activity measured from the peripheral and central auditory systems of young female adults with normal hearing can, in part, account for the large intersubject variability observed in the ANL. In addition to the ABR and MLR, cortical, auditory long latency responses (LLRs) were obtained from two groups of females (one with low ANLs and one with high ANLs). The LLR occurs within 300 ms after stimulus onset and immediately follows the MLR. It is composed of a series of three to four peaks named for their polarity and sequence (P1, N1, P2, and N2), with N1 and P2 as the most stable and reliable of these waves (Näätänen and Picton, 1987). Generators include the primary auditory cortex, auditory association areas, the frontal cortex, and subcortical areas (for review, see Stapells, 2002). By examining AEPs generated from various levels of the auditory system, any differences seen in physiological responses between the two groups will contribute to our understanding of the role the peripheral and central auditory nervous system plays in the acceptance of background noise within a homogeneous population of young female adults.

A secondary purpose of this study was to determine if behavioral and physiological responses from females with low versus high ANLs are differentially affected by the presentation level of the stimuli. Franklin *et al.* (in press) modified the ANL procedure to investigate acceptance of background noise not only for speech at MCL but for a wide range of speech presentation levels. As speech presentation level was changed from 20 to 76 dB HL, the ANLs in-

creased. However, the change in ANL with increasing speech presentation level was not uniform among the listeners. Individuals with low ANLs had a slower rate of growth in ANLs with increases in speech presentation level than individuals with high ANLs.

In the auditory nervous system, low sound presentation levels are coded by a small number of neural responses. Increasing stimulus levels above threshold leads to an increase in neuronal activity; the stronger the stimulus, the greater the number of receptors that are activated (Kandel *et al.*, 2000). Thus, increases in intensity lead to greater neural synchrony and faster conduction times which translates into increased amplitudes and decreased absolute latencies of AEPs. To explore the possibility that stimulus intensity will differentially affect AEPs in females with low versus high ANLs, two tone bursts (500 and 3000 Hz) at two presentation levels (35 and 70 dB nHL) were used to obtain ABRs, MLRs, and LLRs. For comparison, the ANL procedure was modified to obtain ANLs with the speech presentation level at 35 and 70 dB HL in addition to MCL.

Differences in electrophysiological activity and/or differential effects of stimulus intensity on ANL or AEP measures from females with low versus high ANLs may reflect differences in excitatory and/or inhibitory processing between the groups. For example, preliminary data (Tampas *et al.*, 2004) suggest that females with low ANLs (i.e., greater background noise acceptance) have smaller amplitude AEP peaks relative to females with high ANLs (i.e., lower background noise acceptance). Such responses in females with low versus high ANLs could be indicative of reduced responsiveness in the afferent auditory nervous system or greater inhibitory activity from the central efferent system, or both. If so, increases in presentation level would be expected to cause less of an increase in AEP amplitudes, less of a decrease in AEP latencies, and a slower rate of growth of ANL in females with low versus high ANLs.

II. METHODS

A. Subjects

Participants consisted of two groups of young, normal-hearing female adults (19–37 years, mean age=24 years). Females were chosen because previous research reports more robust electrophysiological responses in females than males for the proposed measures (Onishi and Davis, 1968; Shucard *et al.*, 1981). One group was made up of 11 participants with low ANLs of 6 dB or less, while the second group contained ten participants with high ANLs of 16 dB or greater, for a total of 21 participants. Each participant met the following criteria: normal hearing (thresholds of 15 dB HL or lower for pure tones at 0.5, 1, 2, 3, 4, 6, and 8 kHz in both ears) as determined by audiometric testing; normal middle ear function as determined by immitance testing (i.e., type A tympanograms); right handed to minimize possible hemispheric effects; and no known otological, neurological, or learning deficits as reported by subjects. Participants not on oral contraception were tested one week prior to menses because this time has been shown to have the least effect on AEPs (Elkind-Hirsch *et al.*, 1992).

B. Interview-screening procedure

Initially, all participants were interviewed and screened before continuing with the experiment. First, a description of the study, including the general purpose, nature of participation, and risks and benefits was given. Then, an informed consent statement was read and signed and a case history was given, asking specific questions about past and present ear infections, other otologic diseases, and neurological or learning deficits. A hearing evaluation, including audiometric testing, tympanometry, and otoscopy was conducted on each participant and ANL at MCL was measured to determine qualification for group assignment (See Acceptable noise level procedure). Participants with ANLs that did not fall within the range of the two defined groups were excluded from the study. At the completion of the interview-screening procedure, the qualifying participants continued on with physiological testing.

C. Acceptable noise level procedure

The test materials and procedure for establishing ANL in the interview-screening session were based on the Nabelek *et al.* (2004) methods. Acceptance of background noise was measured in a sound-treated booth with a recording of running speech with a male voice (Arizona Travelogue, Cosmos, Inc., Kelowna, British Columbia) as the primary stimulus against an eight person multitalker babble as the competing stimulus. Three speech presentation level conditions (35 dB HL, MCL, and 70 dB HL) were presented in random order. The stimuli were presented from a compact disc to a Grason Stadler (GSI-61) audiometer and delivered binaurally via insert earphones (Etymotic ER-3A).

To establish ANL for the MCL condition, the listener was given two hand-held buttons to signal the examiner to adjust the volume of the speech up or down. The verbal and written instructions for measuring MCL were as follows:

> *You will listen to a story through earphones. After a few moments, select the loudness of the story that is most comfortable for you, as if listening to a radio. Hand-held buttons will allow you to make adjustments. First, turn the loudness up until it is too loud and then down until it is too soft. Finally, select the loudness level that is most comfortable for you.*

The loudness level of the running speech began at 0 dB HL and was adjusted in steps of 10 dB when establishing the maximum and minimum loudness levels. Once the listener indicated the story was "too soft," the level of the story was adjusted in 2-dB increments.

To establish BNL, the speech was presented at MCL, 35 dB HL, or 70 dB HL, and then the background noise was added. As with the MCL procedure, the listener was given two hand-held buttons to signal the examiner to adjust the volume of the background noise up or down. The verbal and written instructions for measuring BNL in reference to these three speech presentation levels were as follows:

> *You will listen to the same story with background noise of several people talking at the same time. After you have listened to this for a few moments select the maximum level of the background noise*

that you would be willing to accept or "put-up-with" without becoming tense and tired while following the story. First, turn the noise up until it is too loud and then down until the story becomes very clear. Finally, adjust the noise (up and down) to the level that you would "put-up-with" for a long time while following the story.

The loudness level of the noise began at 0 dB HL and was adjusted in steps of 10 dB when establishing the maximum and minimum loudness levels. Once the listener indicated the story was "very clear," the level of the noise was adjusted in 2-dB increments.

The ANL was calculated by subtracting the BNL from the speech presentation level (e.g., ANL=MCL-BNL). The ANL for each speech presentation level was repeated three times and the average of the three was used. The ANL performed at the MCL was used in determining group qualification.

D. Physiological procedure

Physiological measures included ABRs, MLRs, and LLRs. Participants were seated in a sound-treated booth, comfortably reclined in an armchair with their heads and necks well supported, while they watched a silent videotaped movie with closed captioning and did not sleep. The video display served to maintain subject alertness and reduce random electro-oculographic (EOG) interference. The order of the AEP acquisition was randomized across subjects.

The AEPs were recorded to 500 and 3000 Hz Blackman-gated tone-burst stimuli of negative polarity generated with a 16-bit digital-to-analog converter (Tucker-Davis Technologies, Alachua, FL, model DA1). The gated tone bursts had rise/fall times of two cycles and a plateau of one cycle (Davis *et al.*, 1984; Gorga and Thornton, 1989). The tone bursts were presented at a rate of 1.1/s and levels of 35 dB nHL (70 dB peak SPL) and 70 dB nHL (104 dB peak SPL), for recording "low" and "high" AEPs, respectively. Weekly calibrations with a sound level meter (Bruel and Kjaer, type 2230) set to peak hold capability and a 2-cc coupler (Zwislocki type) confirmed the accuracy of the signal level, which should be kept within 1 dB peak SPL. The stimuli were delivered diotically via electrically shielded insert earphones (Etymotic Research, Elk Grove Village, IL, model ER-3A).

AEPs were acquired with a four-channel electrode configuration using gold-plated electrodes applied to the surface of the scalp and held in place with medical tape. Electrode impedances were measured at 30 Hz, were below 5 k Ω, and within 1 k Ω of each other. The non-inverting electrodes were placed along the center of the head (C_z, C_5, C_6), with linked reference electrodes on both ears and the ground electrode on the forehead (Fpz) (Jasper, 1958). Among participants, the noninverting electrodes were randomly assigned to channels one, two, and three. Using the fourth electrode channel, the EOG was measured to develop an eye-blink rejection rule for each subject. The EOG was recorded between electrodes above and below one eye and amplified (gain: 1×10^4). Artifact rejection, used to exclude samples collected during eye blinks or other muscular contractions, was set at a level equivalent to that of the smallest recorded EOG during a series of ten blinks performed just prior to data collection. Accordingly, an artifact rejection algorithm was applied to the on-line averaging waveform. If the peak voltage within a sweep exceeded ±80 μV, that sweep was excluded from the averaged waveform. The ongoing average of the AEP waveform was monitored. Tucker-Davis Technologies (Alachua, FL) SIGGEN® and BIOSIG® software was used for data acquisition. The AEPs were differentially amplified (gain: 2×10^5) (Tucker-Davis Technologies, model DB4) and filtered. The rejection rate of these filters was -6 dB/octave and the bandwidth was set at 10–3000 Hz for the ABRs and MLRs, and at DC–30 Hz for the LLRs. The time window for the ABRs and MLRs was 145 ms, which included a 70-ms prestimulus period. The time window for the LLRs was 750 ms, which included a 350-ms prestimulus period. Each response was digitized via a 16-bit analog-to-digital converter (Tucker-Davis Technologies, model AD1). A total of 1000 sweeps were presented for ABR and MLR recordings. For LLRs, 250 sweeps were presented at each recording. The sampling rate was set at 10 kHz for all AEPs.

For all stimuli, the ABR and MLR waveforms were obtained simultaneously. One ABR/MLR and one LLR waveform were obtained from each of three channels, unless a replication was needed for identification of particular components. AEPs were evoked to two stimulus levels (35 dB nHL and 70 dB nHL) at two different tone-burst frequencies (500 and 3000 Hz). Thus, for ABR and MLR recordings, 12 waveforms (3 electrodes $\times$ 2 levels $\times$ 2 tone-burst stimuli) per subject acquired; For LLR recordings, there were also 12 waveforms (3 electrodes $\times$ 2 levels $\times$ 2 tone-burst stimuli) per subject were acquired. Overall, a total of 24 waveforms were obtained for each subject.

Each AEP waveform was analyzed using Tucker-Davis Technologies software (BIOSIG®). Peaks were selected based on latencies, which were expected to fall within a certain time frame. ABR amplitude (peak to following trough) and absolute latency was measured for waves I, III, and V. MLR amplitude (peak-to-peak) was measured between waves Na and Pa, and absolute latency was measured for waves Na and Pa. LLR amplitude (peak-to-peak) was measured between waves P1 and N1 and N1 and P2. The absolute latency was measured for waves P1, N1, and P2. If a waveform was not present, the peak-to-peak amplitude was recorded as 0 μV and the absence of a latency was noted (e.g., Kavanagh *et al.*, 1984). For the purpose of presentation and for comparison with the results of other studies, AEP response latencies were corrected for delays imposed during data acquisition. The length of the tubes for the insert earphone causes a 0.9-ms delay and the amplifier causes a 2-ms group delay. Thus, 2.9 ms in addition to the time between sweep onset and stimulus onset for a given set of AEPs were subtracted from the raw latency values. The data presented are from the Cz electrode. [Note: The data from C5 and C6 are currently being evaluated and will be used to investigate hemispheric differences among groups with high and low acceptance of background noise when listening to speech, a topic not addressed in the current manuscript.]

TABLE I. Individual and group data of acceptable noise levels (ANL) when listening to speech for 35 dB HL, MCL, and 70 dB HL speech presentation levels in females with high and low ANLs.

		Speech presentation level (dB HL)		
	Participant	35	MCL	70
High ANL group	S1	9.0	17	23
(***N*=10**)	S3	16.0	17	30
	S4	17.5	27	28
	S5	18.0	22	35
	S6	10.0	22	28
	S8	3.5	17	28
	S9	15.5	21	31
	S10	14.0	21	29
	S18	4.5	21	20
	S19	9.0	20	27
	Mean	**11.7**	**20.5**	**27.9**
	Standard deviation	**4.9**	**3.1**	**3.8**
Low ANL group	S2	2	3	5
(***N*=11**)	S7	1	0	9
	S11	3.5	2	15
	S12	−1	2	2
	S13	5	4	10
	S14	1.5	3	5
	S15	0	−3	2
	S16	2	3	6
	S17	1	3	5
	S20	1	5	9
	S21	−2	1	5
	Mean	**1.3**	**1.9**	**6.6**
	Standard deviation	**1.9**	**2.2**	**3.8**

III. RESULTS

A. Acceptable noise levels

Individual and group ANL data are reported in Table I. When obtained at MCL, the mean ANL was 20.5 and 1.9 dB for the high and low ANL groups, respectively. Average MCLs were the same for each group (53 dB HL) with a standard deviation (S.D.) of 2.9 for the high ANL and 3.6 for the low ANL group. Average BNLs were 32 dB HL (S.D. =5.9) and 51 dB HL (S.D.=4.5) for the high and low ANL groups, respectively. Acceptable noise levels for the 35 and 70 dB speech presentation levels were 11.7 and 27.9 dB for the high ANL group and 1.3 and 6.6 dB for the low ANL group, respectively.

A two-factor, repeated measures analysis of variance (ANOVA) was conducted on ANL. The factors were group (two levels: high ANL, low ANL) and speech presentation level (repeated measures on three levels, 35 dB HL, MCL, and 70 dB HL). Significant main effects were found for group
($F_{1,19}=181.88$, $p<0.001$) and speech presentation level ($F_{2,18}=109.13$, $p<0.001$). The group by speech presentation level interaction was significant ($F_{2,18}=32.35$, $p<0.001$), such that females from the low ANL group had less of an increase in ANL with increasing speech presentation level versus females from the high ANL group (Fig. 1). Rate of ANL growth as a function of speech presentation was calculated to be 0.15 dB/dB for females with low ANLs and 0.44 dB/dB for females with high ANLs.

B. Auditory evoked potentials

The independent variables of greatest interest were group (high versus low ANL), stimulus level (35 dB nHL versus 70 dB nHL), and stimulus frequency (500 Hz versus 3000 Hz). All components of the AEP response were present for all listeners to all stimulus types, except for wave I of the ABR to the low-level, 500- and 3000-Hz tone bursts. Wave I was expected to disappear into the noise floor at such low stimulus levels (Hall, 1992). Mean data comparing ABR, MLR, and LLR waves of the two groups for the 35 dB nHL and 70 dB nHL 500- and 3000-Hz stimuli are presented in Tables II and III, respectively. To underscore group differences in individual AEP components, amplitude data collapsed across frequency and level are presented in Fig. 2.

To examine differences in AEP responses between females with low versus high ANLs, four repeated measures, three-factor multivariate ANOVAs (MANOVAs) were conducted; one on ABR peak-to-peak amplitudes (waves I, III, and V), one on ABR latencies (waves III and V), one on MLR and LLR amplitudes (Na-Pa, P1-N1, N1-P2), and one on MLR and LLR latencies (Na, Pa, P1, N1, P2). The factors were ANL group (two levels, high versus low), frequency (repeated measures on two levels, 500 Hz versus 3000 Hz), and stimulus presentation level (repeated measures on two levels, 35 dB nHL versus 70 dB nHL). Post-hoc ANOVAs were used to follow up on all main effects or interactions. The main effect of group and the interaction effects between group and stimulus level or between group and stimulus fre-

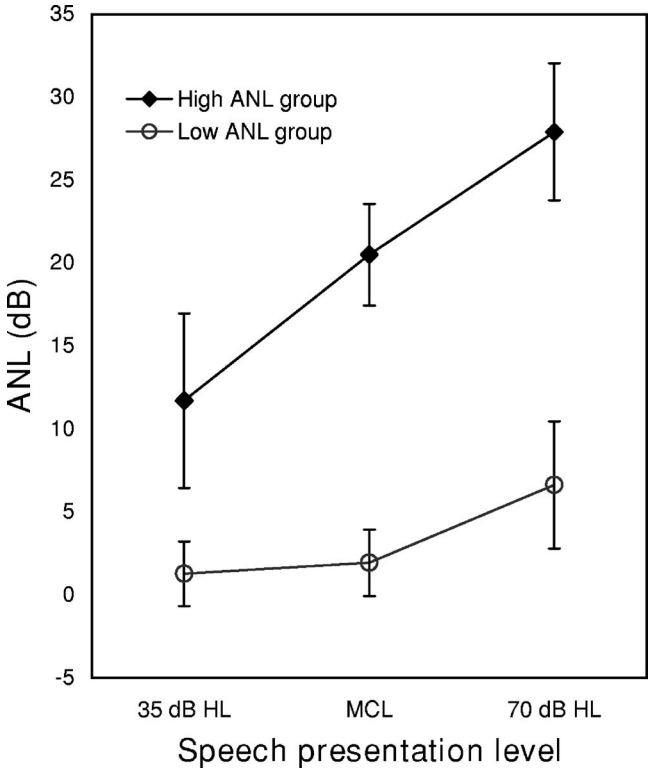


FIG. 1. Mean acceptable noise levels (ANL) while listening to speech for high ($n=10$) and low ($n=11$) ANL groups as a function of speech presentation level.

quency are the main focus of the results section. Other main effects and interactions are mentioned only briefly.

Additionally, because wave I latency was typically absent to low-level stimuli and could not be included in the MANOVA, a two-factor repeated measures ANOVA was conducted on wave I latencies obtained with high-level stimuli. The factors were ANL group (two levels) and stimulus frequency (repeated measures on two levels). It should be noted that because absent wave I peak-to-peak amplitudes were recorded as zero (Kavanagh *et al.*, 1984), wave I amplitude could be included in the overall MANOVA on ABR and MLR amplitudes. However, to be sure that labeling absent wave I amplitudes as zero did not skew the data analysis, a two-way ANOVA also was conducted on wave I amplitude and a three-way MANOVA on waves III, V, and Na -Pa amplitudes. Doing so did not alter any statistical findings and so these tests are not reported.

1. ABR amplitude

The three factor MANOVA on ABR waves I, III, and V amplitude revealed no main effect for group ($F_{3,17}=0.84$, $p=0.49$). Significant main effects were found for frequency ($F_{3,17}=3.78$, $p=0.03$) and level ($F_{3,17}=37.32$, $p<0.001$). The interaction between frequency and level also was significant ($F_{3,17}=3.76$, $p=0.031$). No other interactions were significant. Post-hoc univariate testing revealed that the amplitudes of waves I ($F_{1,19}=7.96$, $p=0.011$), III ($F_{1,19}=8.93$, $p=0.008$), and V ($F_{1,19}=6.97$, $p=0.016$) all contributed to the significant main effect of frequency reported by the MANOVA. Examination of the mean data indicated that higher-frequency stimuli evoked smaller amplitude waveforms. Univariate testing also revealed that the amplitudes of waves I ($F_{1,19}=98.2$, $p<0.001$), III ($F_{1,19}=24.1$,$p<0.001$), and V ($F_{1,19}=46.8$, $p<0.001$) all contributed to the significant main effect of level reported by the MANOVA. Examination of the mean data indicated, as expected, that as stimulus level decreased, the amplitudes also decreased. Additionally, the amplitudes of waves I ($F_{1,19}=7.96$, $p=0.011$), III ($F_{1,19}=6.7$, $p=0.018$), and V ($F_{1,19}=7.44$, $p=0.013$) all contributed to the significant frequency by level interaction reported by the MANOVA. The significant interaction indicated that response amplitudes decreased more with a decrease in stimulus level using 500-Hz stimuli versus 3000-Hz stimuli.

2. ABR latency

The two-factor repeated measures ANOVA conducted on wave I latency to high-level stimuli revealed that the main effect of group was not significant ($F_{1,19}=2.09$, $p=0.165$) nor was the interaction. The main effect of frequency was significant ($F_{1,19}=11.59$, $p=0.003$), indicating that, as expected, wave I latencies were longer to lower-frequency stimuli.

The three-factor repeated-measures MANOVA conducted on the latencies of ABR waves III and V indicated a significant main effect for group ($F_{2,16}=9.06$, $p=0.002$), frequency ($F_{2,16}=52.56$, $p<0.001$), and level ($F_{2,16}=52.56$, $p<0.001$). The level by group ($F_{2,16}=3.85$, $p=0.043$), level by frequency ($F_{2,16}=23.56$, $p<0.001$), and group by level by frequency ($F_{2,16}=4.03$, $p=0.038$) interactions were also significant. Individual ANOVAs used to follow up on the main effects revealed that latencies of waves III ($F_{1,17}=8.42$, $p=0.010$) and V ($F_{1,17}=18.22$, $p=0.001$) contributed to the significant main effect of group. Females in the low ANL group tended to exhibit longer absolute latencies than those from the high ANL group. Neither the latencies of waves III ($F_{1,17}=0.49$,$p=0.526$) or V ($F_{1,17}=2.44$, $p=0.136$) contributed individually to the significant level by group interaction reported by the MANOVA. Examining the combined latencies of waves III and V, females with low ANLs tended to have less of a decrease in latency with increasing level than females with high ANLs. Last, univariate tests revealed that wave III ($F_{1,17}=8.52$, $p=0.010$) was responsible for the group by level by frequency interaction reported in the MANOVA. Specifically, with an increase in stimulus level, females with low ANLs had a smaller decrease in latency for high-frequency stimuli and a larger decrease in latency for low-frequency stimuli. In contrast, with an increase in stimulus level, females with high ANLs had a greater decrease in latency for high-frequency stimuli and a smaller decrease in latency for low-frequency stimuli (Fig. 3). Thus, wave III latencies to 500-Hz tone bursts for the two groups were affected similarly with a change in stimulus intensity while wave III latencies for 3000-Hz tone bursts decreased less with an increase in intensity for females with low ANLs versus females with high ANLs.

Additionally, post-hoc ANOVAs revealed that latencies of waves III ($F_{1,17}=108.25$, $p<0.001$) and V ($F_{1,17}=64.65$, $p<0.001$) contributed to the significant main effect of frequency found in the MANOVA. As stimulus frequency

TABLE II. Mean AEP peak latencies (ms) and peak-to-peak amplitudes (μV) measured from Cz in response to 500-Hz tone bursts at 35 and 70 dB nHL from females with high and low ANLs. Standard deviations are in parentheses.

Latencies of AEP components		**35 dB nHL, 500 Hz**	**70 dB nHL, 500** Hz
Wave	Group	Latency (ms)	Latency (ms)
I	High	NP	2.4 (0.60)
	Low	NP	2.6 (0.86)
III	High	6.4 (0.61)	4.5 (0.31)
	Low	7.1 (0.65)	4.8 (0.28)
V	High	8.3 (0.68)	6.6 (0.40)
	Low	8.9 (0.40)	6.8 (0.41)
Na	High	19.7 (1.20)	16.4 (2.70)
	Low	19.9 (1.70)	17.3 (1.20)
Pa	High	30.1 (3.80)	29.0 (2.50)
	Low	30.8 (4.00)	29.9 (4.10)
P1	High	62.8 (11.90)	55.8 (14.80)
	Low	66.0 (8.20)	58.1 (9.00)
N1	High	101.9 (12.80)	88.3 (9.60)
	Low	98.1 (10.00)	88.0 (10.30)
P2	High	152.0 (12.80)	145.2 (8.40)
	Low	137.6 (8.70)	138.9 (6.50)
Amplitudes of AEP components		**35 dB nHL, 500 Hz**	**70 dB nHL, 500** Hz
Wave	Group	Amplitude (μV)	Amplitude (μV)
I	High	0.00 (0.00)	0.17 (0.08)
	Low	0.00 (0.00)	0.18 (0.16)
III	High	0.10 (0.08)	0.23 (0.06)
	Low	0.08 (0.06)	0.28 (0.26)
V	High	0.46 (0.16)	0.83 (0.40)
	Low	0.39 (0.11)	0.72 (0.23)
Na-Pa	High	1.01 (0.41)	1.22 (0.59)
	Low	0.70 (0.20)	0.83 (0.26)
P1-N1	High	1.05 (0.18)	2.13 (1.10)
	Low	0.82 (0.24)	1.06 (0.37)
N1-P2	High	1.55 (0.50)	2.37 (1.30)
	Low	0.99 (0.27)	1.56 (0.38)

Note: NP= wave not present.

increased, the mean latencies decreased. Post-hoc testing also revealed that the latencies of waves III ($F_{1,17}=170.72$, $p<0.001$) and V ($F_{1,17}=487.87$, $p<0.001$) contributed to the significant main effect of level reported in the MANOVA. Examination of the mean data indicated, as expected, that as stimulus level decreased, the latencies increased. The latencies of waves III ($F_{1,17}=39.14$, $p<0.001$) and V ($F_{1,17}=34.02$, $p<0.001$) contributed to the significant level by frequency interaction reported by the MANOVA. The interaction indicated that response latencies decreased more with an increase in level to the 500-Hz stimuli than to the 3000-Hz stimuli.

3. MLR and LLR amplitude

The MANOVA on MLR and LLR peak-to-peak amplitudes revealed significant main effects for group ($F_{3,17}=7.17$, $p=0.003$), frequency ($F_{3,17}=18.44$, $p<0.001$), and level ($F_{3,17}=7.31$, $p=0.002$), as well as a level by group interaction ($F_{3,17}=3.62$, $p=0.035$). Post-hoc univariate tests revealed that Na-Pa ($F_{1,19}=4.67$, $p=0.004$), P1-N1 ($F_{1,19}=15.68$, $p<0.001$), and N1-P2 ($F_{1,19}=8.02$, $p<0.001$) peak-to-peak amplitudes contributed to the significant main effect of group reported in the MANOVA. Peak-to-peak amplitudes were smaller for the females in the low ANL group than for those in the high ANL group. Representative ABR/MLR and LLR tracings from two participants (one with a low ANL and one with a high ANL) are displayed in Figs. 4 and 5, respectively.

P1-N1 ($F_{1,19}=7.67$, $p=0.012$) contributed to the significant level by group interaction seen in the MANOVA. P1-N1 amplitudes for the low ANL group increased less with an increase in stimulus level than those for the high ANL group (Fig. 6).

Additionally, post-hoc univariate tests revealed that Na-Pa ($F_{1,19}=10.68$, $p=0.004$), P1-N1 ($F_{1,19}=25.19$, $p<0.001$), and N1-P2 ($F_{1,19}=4.89$, $p<0.001$) peak-to-peak amplitudes contributed to the significant main effect of frequency reported in the MANOVA. Amplitudes were larger for the 500-Hz stimuli versus the 3000-Hz stimuli. For the main effect of level, P1-N1 ($F_{1,19}=15.76$, $p=0.001$) and N1-P2 ($F_{1,19}=16.97$, $p=0.001$) were significant such that amplitudes increased with an increase in stimulus level.

TABLE III. Mean AEP peak latencies (ms) and peak-to-peak amplitudes (μV) measured from Cz in response to 3000-Hz tone bursts at 35 and 70 dB nHL from females with high and low ANLs. Standard deviations are in parentheses.

Latencies of AEP components		**35 dB nHL, 3000 Hz**	**70 dB nHL, 3000** Hz
Wave	ANL group	Latency (ms)	Latency (ms)
I	High	NP	1.6 (0.60)
	Low	NP	2.0 (0.04)
III	High	5.0 (0.70)	3.8 (0.40)
	Low	4.7 (0.30)	4.1 (0.20)
V	High	6.8 (0.60)	5.8 (0.20)
	Low	7.0 (0.50)	6.1 (0.20)
Na	High	17.0 (2.60)	16.7 (1.10)
	Low	18.6 (0.60)	17.3 (1.80)
Pa	High	26.7 (1.30)	25.4 (1.80)
	Low	25.4 (3.10)	25.3 (2.00)
P1	High	73.6 (14.10)	48.5 (22.20)
	Low	69.0 (18.10)	48.0 (14.90)
N1	High	104.3 (9.30)	86.6 (13.00)
	Low	101.2 (9.10)	89.0 (12.60)
P2	High	158.8 (13.40)	142.0 (8.90)
	Low	150.4 (17.60)	145.0 (15.20)
Amplitudes of AEP components		**35 dB nHL, 3000 Hz**	**70 dB nHL, 3000** Hz
Wave	ANL group	Amplitude(μV)	Amplitude (μV)
I	High	0.00 (0.00)	0.10 (0.03)
	Low	0.00 (0.00)	0.10 (0.03)
III	High	0.09 (0.08)	0.11 (0.06)
	Low	0.05 (0.08)	0.10 (0.10)
V	High	0.57 (0.33)	0.63 (0.40)
	Low	0.32 (0.12)	0.49 (0.10)
Na-Pa	High	1.0 (0.60)	1.15 (0.70)
	Low	0.7 (0.20)	0.73 (0.30)
P1-N1	High	0.7 (0.20)	1.70 (1.00)
	Low	0.6 (0.10)	0.70 (0.20)
N1-P2	High	0.82 (0.31)	1.81 (0.90)
	Low	0.72 (0.22)	1.22 (0.50)

Note: NP= wave not present.

4. MLR and LLR latency

The MANOVA conducted on MLR and LLR latencies revealed a significant main effect for frequency ($F_{5,15}=5.78$, $p=0.004$) and level ($F_{5,15}=8.27$, $p=0.001$) and an interaction for frequency by level ($F_{5,15}=5.06$, $p=0.006$). The main effect of group was not significant ($F_{5,15}=1.101$, $p=0.400$), nor were any other interactions. Subsequent univariate tests revealed that Na ($F_{1,19}=10.41$, $p=0.004$), Pa ($F_{1,19}=25.21$, $p<0.001$), and P2 ($F_{1,19}=4.63$, $p=0.045$) latencies contributed to the significant main effect of frequency found in the MANOVA testing. Similar to ABR latencies, examination of the means revealed that as frequency increased, the mean latencies decreased. Na ($F_{1,19}=23.16$, $p<0.001$), P1 ($F_{1,19}=23.69$, $p<0.001$), N1 ($F_{1,19}=28.17$, $p<0.001$), and P2 ($F_{1,19}=9.81$, $p=0.005$) latencies contributed to the significant main effect of level found in the MANOVA testing. As expected, latencies were longer for lower-level stimuli than for higher-level stimuli. MLR wave Na ($F_{1,19}=7.53$, $p=0.013$), LLR waves P1 ($F_{1,19}=11.15$, $p=0.003$) and P2 ($F_{1,19}=5.27$, $p=0.033$) contributed to the significant interaction between frequency and level such that latencies decreased less with increasing intensity for the 500-Hz stimuli versus the 3000-Hz stimuli.

5. AEP data reduction

To underscore group differences in individual AEP components, responses for each frequency and presentation level were grand averaged to create composite peak-to-peak amplitudes (Fig. 2) and absolute latencies for each ABR, MLR, and LLR wave, respectively. A one-factor, repeated measures MANOVA was conducted on AEP peak-to-peak amplitudes (ABR waves III and V, MLR waves Na-Pa, and LLR waves P1-N1 and N1-P2) and absolute latencies (ABR waves III and V, MLR waves Na and Pa, and LLR waves P1, N1, P2). The factor was group (high ANL, low ANL). The MANOVA on peak-to-peak amplitudes revealed significant main effects for group ($F_{6,14}=4.06$, $p=0.016$). Post-hoc univariate tests revealed that Na-Pa ($F_{1,19}=4.67$, $p=0.044$), P1-N1 ($F_{1,19}=15.68$, $p=0.001$), and N1-P2 ($F_{1,19}=8.02$, $p=0.011$) peak-to-peak amplitudes contributed to the significant main effect

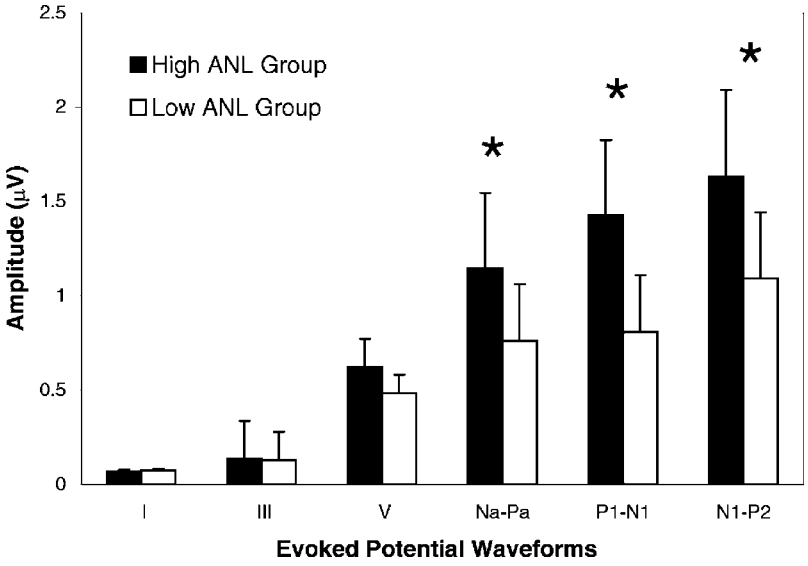


FIG. 2. Evoked potential amplitude data collapsed across frequency and stimulus presentation level for all waveforms. Significant group differences ($p<0.05$) are noted by ★.

of group reported in the MANOVA such that amplitudes were smaller for the females in the low ANL group than for those in the high ANL group.

The MANOVA on absolute latency also revealed a significant main effect for group ($F_{6,14}=3.10$, $p=0.038$). Univariate testing revealed that only ABR wave V ($F_{1,19}=13.25$, $p=0.002$) contributed to the significant main effect of group reported in the MANOVA, such that females in the low ANL group exhibited longer absolute latencies than those in the high ANL group.

IV. DISCUSSION

One purpose of the current study was to investigate whether individual differences in physiological activity measured from the auditory nerve to the cortex of young female adults with normal hearing are related to the large intersubject variability observed in acceptable noise level (ANL). Specifically, ABRs, MLRs, and LLRs representing physiologic activity from the auditory nerve to cortical regions of the auditory system were obtained from females with low and high ANLs. Results indicate that differences in the responsiveness of more central regions of the auditory nervous system do appear to be related to differences in ANL. A central influence on ANL is consistent with findings by Harkrider and Smith (2005) that the amount of background noise listeners were willing to accept when listening to speech in a monotic condition correlated with the amount of background noise they were willing to accept in a dichotic condition. They interpreted this to indicate the ANL response is mediated, at least in part, at a level beyond the superior olivary complex where binaural processing begins.

In the current study, Fig. 2 represents the combined peak-to-peak amplitudes from the "low" and "high" 500-Hz and 3000-Hz tone-burst stimuli. This composite demonstrates that the high ANL group, overall, exhibited larger amplitudes than the low ANL group and reveals a trend developing across the AEP types. ABR waves I and III are comparable between the two groups, but the amplitude differences become increasingly more remarkable in wave V of the ABR as well as the MLR and LLR waves. Considering the structures from which these waves are thought to originate, the data suggest that processes occurring at more central levels of the auditory system are influencing the ANL response. This conclusion is in accord with that of Tampas *et al.* (2004) in which group differences were not seen for OAEs or early waves of the ABR but were displayed in wave V amplitude of the ABR and Na-Pa of the MLR from a small number of females. In contrast to the current data in which ANL group differences were found for ABR waves III and V latency, ABR latency differences were not reported in Tampas *et al.* (2004). This difference may be due to lower statistical power associated with the smaller number of subjects participating in the pilot study versus the current one.

A second purpose of the current study was to investigate whether the presentation level of the stimuli had differential effects on ANLs and AEPs when measured from females with low versus high ANLs. Results indicate that presentation level interacts with ANL group to influence behavioral judgments of background noise acceptance, as well as some of the physiological measures. Overall, for the 35-dB HL speech presentation level, the ANLs decreased in reference to the presentation at MCL; therefore, listeners were more willing to accept noise when the intensity of the speech decreased. For the 70-dB HL speech presentation level, the ANLs increased in reference to the presentation at MCL. In other words, all listeners were less willing to accept noise when the intensity of the speech increased. These findings are in agreement with reports by Franklin *et al.* (in press), who demonstrated that acceptance of background noise is directly related to the presentation level of the speech stimuli. Notably, the interaction between group and speech presentation level reveals that the ability to accept background noise when listening to speech is dramatically affected by an increase in speech presentation level in females with high versus low ANLs. Specifically, as speech presentation level was increased from 35 to 70 dB HL, the rate of ANL growth ranged from 11 to 28 dB (slope=0.44 dB/dB) for the females with high ANLs and from 1 to 6.5 dB (slope=0.15 dB/dB) for the females with low ANLs (Fig. 1). Because MCL was 53 dB HL for both groups, the differ-

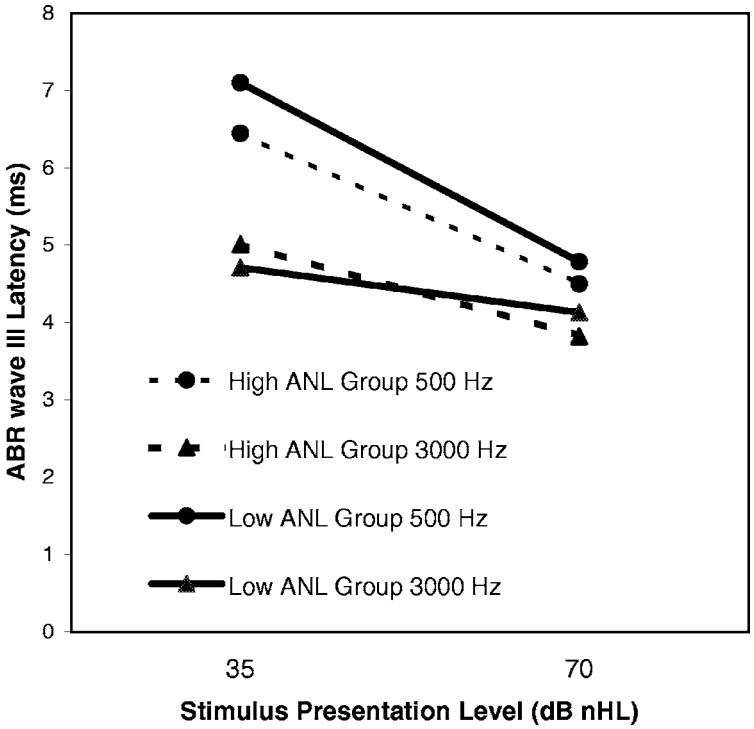


FIG. 3. ABR wave III latency for high and low ANL groups as a function of stimulus presentation level for 500 and 3000 Hz tone bursts.

ence in rate of ANL growth cannot be attributed to differences in overall dynamic range. These results suggest that the auditory processing mechanisms of listeners who are willing to accept large amounts of background noise are distinct from those who are not willing to accept large amounts of background noise. Auditory evoked potential findings from the current study appear to shed light on the gross location of these mechanisms in the auditory system.

The ABR latency differences found for waves III and V suggest that the low ANL group exhibited slower neural conduction times than the high ANL group at the level of the cochlear nucleus, superior olivary complex, and lateral lemniscus. Further, ABR latencies for waves III and V also were differentially affected by the presentation level of the stimuli implying that the nervous systems of the females with low ANLs were less responsive to a change in stimulus intensity than those with high ANLs. These findings, in addition to the fact that females with low ANLs had smaller Na-Pa, P1-N1, and N1-P2 peak-to-peak amplitudes than females with high ANLs suggest that (1) central efferent mechanisms are stronger in the group with low ANLs such that sensory inputs are suppressed more than in the high ANL group and/or (2) that central afferent mechanisms are less active in the low ANL versus the high ANL group. This consideration

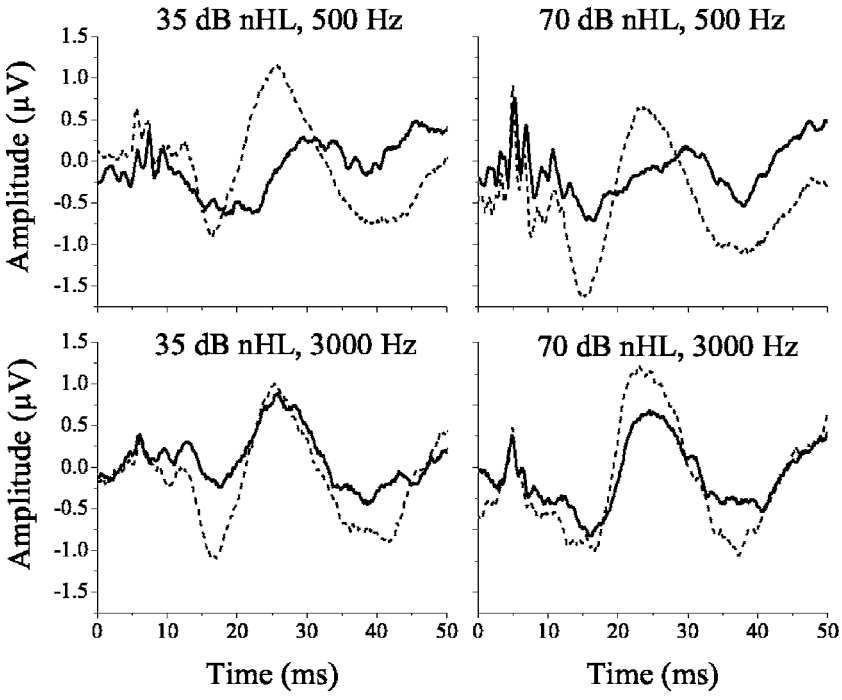


FIG. 4. Representative ABR and MLR waveforms from two participants (one from each ANL group). The solid lines represent the waveforms from the low ANL group and the dashed lines represent the waveforms from the high ANL group. Note: In order to better visualize the ABR and MLR on the same graph, the time window of the responses has been truncated, excluding EEG data from prestimulus and post-MLR time frames. Data are recorded from Cz.

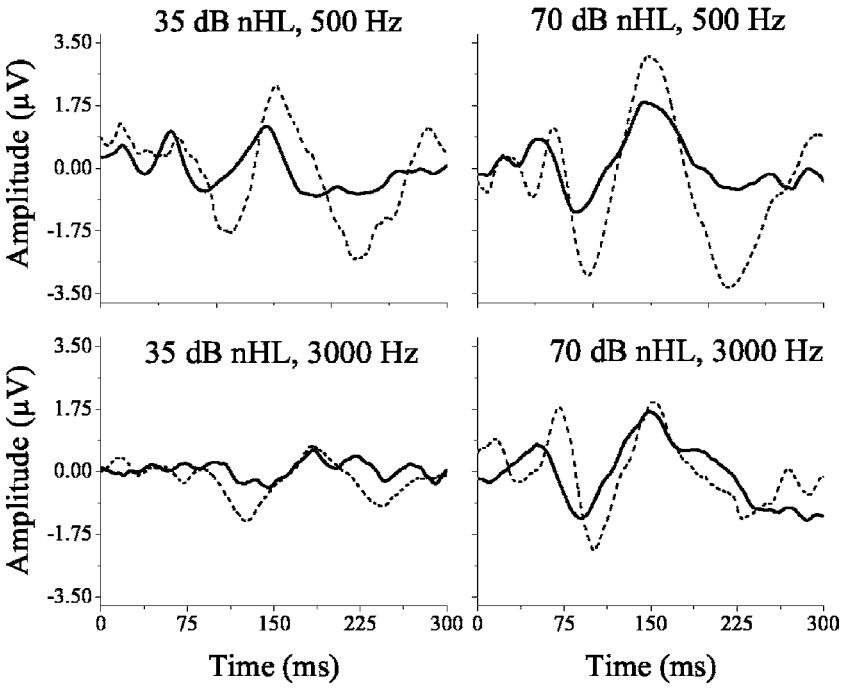


FIG. 5. Representative LLR waveforms from two participants (one from each ANL group). The solid lines represent the waveforms from the low ANL group and the dashed lines represent the waveforms from the high ANL group. Note: In order to better visualize the LLR, the time window of the response has been truncated, excluding EEG data from pre-stimulus and post-LLR time frames. Data are recorded from Cz.

is in agreement with electrophysiologic studies reporting enhanced AEPs to be indicative of increased central nervous system (CNS) excitability and the reduced capacity for inhibition (Ahveninen *et al.*, 1999; Harkrider and Champlin, 2001a, b; Lei-Zhang *et al.*, 2001).

Unlike the visual and somatosensory pathways, the auditory system lacks any direct pathway from peripheral receptors to the cortex or vice versa (Moore, 1994). Rather, information ultimately reaching and returning from the auditory cortex undergoes significant modification as it is processed. Research on the anatomic and chemical composition of the auditory pathway suggests that inhibition plays an extremely important role at all levels of the system in producing the precise responses of central neurons (for review, see Jackler and Brackmann, 1994). Inhibition is also a major component of selective attention and is manifested in the suppression of task-irrelevant stimuli. The process of selective attention is one in which an individual "selectively attends to some stimuli, or aspects of stimuli, in preference to others" (Kahneman, 1973), and is a vital component of one's ability to function in the environment. There are numerous studies emphasizing the role of the cortex in this ability to effectively suppress stimuli that are not relevant to the task at hand (e.g., Glosser and Goodglass, 1990; Knight *et al.*, 1999; Rueckert and Grafman, 1996; Stern *et al.*, 2001). Reviews of literature in cognitive functioning emphasize that individual differences in behavioral performance might best be explained by underlying differences in cortical inhibition (e.g., Dempster, 1992; Harnishfeger and Bjorkland, 1994). Furthermore, electrophysiologic studies interpret enhanced AEP responses to be indicative of increased brain excitability or a reduced capacity for inhibition, relative to controls (Ahveninen *et al.*, 1999; Lei-Zhang *et al.*, 2001; Harkrider and Champlin, 2001a, b). One implication of these complexities of central organization seems to relate to the phenomenon of background noise acceptance when listening to speech. Results from the current study agree with theories and findings related to individual differences in perceptual responses to similar sensory stimulation, and suggest that there are distinct excitatory and inhibitory mechanisms that interact to maintain sensitivity to background noise.

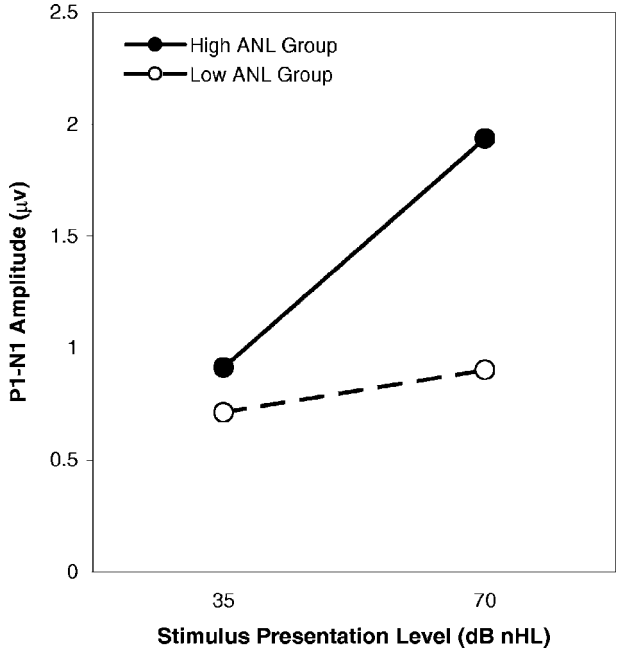


FIG. 6. LLR wave P1-N1 amplitude for high and low ANL groups as a function of stimulus presentation level.

Theories addressing the systematic modulation of sensory input were originally formulated by Petrie (1960, 1967) on the basis of differences observed in the Kinesthetic Figural Aftereffects test (KFA). This test measures the change in perceived tactile width of a block following tactile exposure to a larger block. This concept purported that individuals

differed in the way they reacted to varying intensities of stimuli. Augmenters displayed amplified responses to increasing intensities of incoming stimuli whereas reducers attenuated or showed minimal response enhancement to increased stimuli. Although Petrie pointed out the importance of the augmentation/reduction (AR) phenomenon in understanding interindividual differences in the perception of pain, evoked potential augmentation and reduction has also been observed (Buchsbaum and Silverman, 1968) and seen to reflect variability in cortical processing of different stimulus intensities. Buchsbaum (1976) suggested that the augmentation/reduction (AR) theory is related to central control factors rather than peripheral mechanisms. The application of this neurophysiological model to background noise acceptance in the current study may help explain the differences observed between the high and low ANL groups. Specifically, the high ANL group displayed more of an AEP amplitude enhancement to increasing stimulus intensity than the low ANL group. Similarly, with increasing speech presentation level, the high ANL group exhibited a larger increase in ANL than the low ANL group.

Future research on the identification of variables contributing to differences in background noise acceptance should focus on expanding the subject population and types of responses measured. Further, electrophysiological investigation should be broadened to include males and females of various ages with normal and impaired hearing. The current study considered females with extreme values of performance on the ANL task; however, listeners with more common ANLs (e.g., 10 dB) should be investigated to further understand the physiologic mechanisms influencing ANL. Also, future research should include more cognitive-based electrophysiologic responses. Event-related potentials (ERPs) are brain potentials that are time locked to the occurrence of sensory, motor, or cognitive events and extracted from the ongoing EEG. Auditory ERP recordings provide a window into the brain processes underlying the sensory and perceptual processes of the auditory system and give insight into the physiological/psychological demands of the situation. Addressing these issues would provide more understanding about the physiological and cognitive processes contributing to background noise acceptance.

It is important to point out that this study aimed to characterize the crucial differences in physiological activity exhibited between two subject groups displaying different perceptual patterns of background noise acceptance. The ANL procedure was measured with a running speech and multitalker babble stimulus, while the AEPs were elicited with brief tone bursts. Differences between the stimuli used to obtain ANL and the AEP measures preclude any cause-and-effect conclusions regarding the neural processes contributing to an individual's background noise acceptance. Instead, group differences in electrophysiological responses are best interpreted as indicating that differences in the responsiveness of central regions of the auditory system are related to variability in listeners' willingness to accept background noise. Future studies using more comparable stimuli for the perceptual and physiological stimuli may contribute to the identification of the neural mechanisms responsible for the amount of background noise an individual is willing to accept.

Last, in the current study, group differences were seen for some AEP measures and not others, as well as for amplitudes but not latencies for some of the same AEP components. However, this is not an uncommon finding (e.g., McFadden and Champlin, 2000). Clarifying this will involve more understanding than currently is available regarding the generators of the various AEP waves and the physiological mechanisms involved in mediating acceptance of background noise. There is not a one-to-one association between individual AEP peaks and single generator sites within the brain. Instead, multiple populations at different locations are involved, especially for the later waves (Hall, 1992). Identifying exactly which neural populations are responsible for the group differences observed in the AEPs is beyond the scope or intent of this paper. Further, when using AEPs, the neural generators are primarily auditory ones. It is likely that whatever neural differences are eventually found to mediate acceptance of background noise in the auditory system may be merely physiological indicators of gross neural differences that are responsible for acceptance of stimuli in all sensory modalities.

V. CONCLUSIONS

One goal of this study was to investigate the possibility that differences in judgment of background noise when listening to speech exhibited among individuals may be related to underlying physiological processes. Group differences in electrophysiological measures indicate that differences in the responsiveness of central regions of the auditory system account, at least in part, for variability in listeners' willingness to accept background noise. Differences in AEPs between the two groups were absent for the early ABR waves and emerged for waves III and V of the ABR, as well as for MLR and LLR peaks. Specifically, females with low ANLs had longer wave III and V latencies and smaller Na-Pa, P1-N1, and N1-P2 amplitudes than females with high ANLs.

A second goal of this study was to determine whether the presentation level of the stimuli had differential effects on background noise acceptance and AEPs when measured from females with low versus high ANLs. Results indicate that presentation level interacts with ANL group to influence behavioral judgments of background noise acceptance, as well as some of the physiological measures. Females with low versus high ANLs demonstrated a slower rate of growth in ANL with increasing stimulus presentation level. Additionally, there was a smaller decrease in wave III and V latencies and a smaller increase in N1-P2 amplitudes with increases in stimulus intensity in females with low versus high ANLs.

The main effects of group on the AEP data and the interactions between group and stimulus level on the ANL and some of the AEP data were consistent with the theory that reduced responsiveness of the central afferent auditory nervous system and/or increased strength of cortical inhibition contribute to greater acceptance of background noise (lower

ANLs). Specifically, females with low versus high ANLs demonstrated only minimal growth in ANL with increasing speech presentation level, a behavioral finding in accord with the physiological findings that, for some AEP measures, females with low versus high ANLs demonstrated less of an increase in amplitude or a decrease in latency with increasing stimulus intensity.

ACKNOWLEDGMENTS

The authors thank Dr. Anna K. Nabelek, Dr. Patrick N. Plyler, Dr. Samuel B. Burchfield, Dr. James W. Thelin, and the reviewers for their time and assistance with this project and manuscript preparation.

Ahveninen, J., Jaaskelainen, I. P., Pekkonen, E., Hallberg, A., Hietanen, M., Näätänen, R., and Sillanaukee, P. (**1999**). "Post-withdrawl changes in middle-latency auditory evoked potentials in abstinent human alcoholics," Neurosci. Lett. **268**, 57–60.

Ashmore, J. F. (**1991**). "The electrophysiology of hair cells," Annu. Rev. Physiol. **53**, 465–476.

Buchsbaum, M. (**1976**). "Self-regulation of stimulus intensity: Augmenting/reducing and the average evoked response," in *Consciousness and Self-Regulation*, edited by G. E. Schwartz and D. Shapiro (Plenum, New York), pp. 101, 135.

Buchsbaum, M., and Silverman, J. (**1968**). "Stimulus intensity control and the cortical evoked response," Psychosom. Med. **30**, 12–22.

Chao, L. L., and Knight, R. T. (**1997**). "Prefrontal deficits in attention and inhibitory control with aging," Cereb. Cortex **7**, 63–69.

Crowley, H. J., and Nabelek, I. V. (**1996**). "Estimation of client-assessed hearing aid performance based upon unaided variables," J. Speech Hear. Res. **39**, 19–27.

Davis, H. (**1983**). "An active process in cochlear mechanics," Hear. Res. **9**, 79–90.

Davis, H., Hirsh, S. K., Popelka, G. R., and Formby, C. (**1984**). "Frequency selectivity and thresholds of brief stimuli suitable for electric response audiometry," Audiology **23**, 9–74.

Dempster, F. N. (**1992**). "The rise and fall of the inhibitory mechanism: Toward a unified theory of cognitive development and aging," Dev. Rev. **12**, 45–75.

Dustman, R. E., Emmerson, R. Y., and Shearer, D. E. (**1996**). "Life span changes in electrophysiological measures of inhibition," Brain Cogn **30**, 109–126.

Elkind-Hirsch, K. E., Stoner, W. R., Stach, B. A., and Jerger, J. (**1992**). "Estrogen influences auditory brainstem responses during the normal menstrual cycle," Hear. Res. **60**, 143–148.

Franklin, C. A., Thelin, J. W., Nabelek, A. K., and Burchfield, S. B. "The effect of speech presentation level on acceptance of background noise in normal-hearing listeners," J. Am. Acad. Audiol (in press).

Freyaldenhoven, M. F., Thelin, J. W., Plyler, P. N., Nabelek, A. K., and Burchfield, S. B. (**2005**). "Effect of stimulant medication on the acceptance of background noise in individuals with attention deficit/hyperactivity disorder," J. Am. Acad. Audiol **16**(9), 677–686.

Glosser, G., and Goodglass, H. (**1990**). "Disorders in executive control functions among apahasics and other brain-damaged patients," J. Clin. Exp. Neuropsychol **12**, 485–501.

Gorga, M. P., and Thornton, A. R. (**1989**). "The choice of stimuli for ABR measurements," Ear Hear. **10**, 217–230.

Hall, J. (**1992**). *Handbook of Auditory Evoked Responses*, (Allyn and Bacon, Boston), pp. 50–55, 129–134, 45–49.

Harkrider, A. W., and Champlin, C. A. (**2001a**). "Acute effect of nicotine on non-smokers: II. MLRs and 40-Hz responses," Hear. Res. **160**, 89–98.

Harkrider, A. W., and Champlin, C. A. (**2001b**). "Acute effect of nicotine on non-smokers: III. LLRs and EEGs," Hear. Res. **160**, 99–110.

Harkrider, A. W., and Smith, B. (**2005**). "Acceptable noise level, phoneme recognition in noise, and auditory efferent measures," J. Am. Acad. Audiol **16**, 530–545.

Harnishfeger, K. K., and Bjorkland, D. F. (**1994**). "A developmental perspective on individual differences in inhibition," Learn. Individ. Differ., **6**, 331–355.

Hashimoto, I. (**1982**). "Auditory evoked potentials recorded directly from the human VIIIth nerve and brain stem: Origins of their fast and slow components," Electroencephalogr Clin. Neurophysiol. Suppl. **36**, 305–314.

Jackler, R. K., and Brackmann, D. E. (**1994**). *Neurotology* (Mosby Publishers, St. Louis, MO).

Jacobson, G. P., Privitera, M., Neils, J., Grayson, A., and Yeh, H. (**1990**). "The effects of anterior temporal lobectomy on the middle-latency auditory evoked potential," Electroencephalogr Clin. Neurophysiol. Suppl. **75**, 230–241.

Jasper, H. H. (**1958**). "Report of the committee on methods of clinical examination in electro-encephalography. The ten-twenty electrode system," Electroencephalogr Clin. Neurophysiol. Suppl. **10**, 371–375.

Jerger, J., Oliver, T., and Chmiel, R. (**1988**). "Auditory middle latency response: a perspective," Semin. Hear. **9**, 75–86.

Kahneman, D. (**1973**). *Attention and Effort* (Prentice-Hall, Englewood Cliffs, N.J.), p. 3.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (**2000**). *Principles of Neural Science*, 4th ed. (McGraw-Hill, New York), pp. 123–134.

Kavanagh, K. T., Harker, L. A., and Tyler, R. S. (**1984**). "Auditory brainstem and middle latency responses. I: Effects of response filtering and waveform identification. II: Threshold responses to a 500-Hz tone pip," Ann. Otol. Rhinol. Laryngol. **93**, 1–12.

Knight, R. T., Staines, W. R., Swick, D., and Chao, L. L. (**1999**). "Prefrontal cortex regulates inhibition and excitation in distributed neural networks," Acta Psychol. **101**, 158–178.

Kujawa, S. G., and Liberman, M. C. (**1997**). "Conditioning-related protection from acoustic injury: Effects of chronic de-efferentation and sham surgery," J. Neurophysiol. **78**, 3095–3106.

Lei-Zhang, X., Cohen, H. L., Porjesz, B., and Begleiter, H. (**2001**). "Mismatch negativity in subjects at high risk for alcoholism," Alcohol Clin. Exp. Res. **3**, 230–337.

McDowd, J. M., and Filion, D. L. (**1992**). "Aging, selective attention, and inhibitory processes: A psychophysiological approach," Psychol. Aging **7**, 65–71.

McFadden, D., and Champlin, C. A. (**2000**). "Comparison of auditory evoked potentials in heterosexual, homosexual, and bisexual males and females," J. Assoc. Res. Otolaryngol. **1**(1), 89–99.

Moller, A. (**1994**)."Neural generators of auditory evoked potentials," in *Principles and Applications in Auditory Evoked Potentials*, edited by J. Jacobson (Allyn and Bacon, Boston, MA). pp. 23–46.

Moore, J. K. "The human brainstem auditory pathway," in *Neurotology*, edited by R. K. Jackler and D. E. Brackmann (Mosby Publishers, St. Louis, MO, **1994**). pp. 1–18.

Näätänen, R., and Picton, T. (**1987**). "The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure," Psychophysiology **24**, 375–425.

Nabelek, A. K., Tampas, J. W., and Burchfield, S. B. (**2004**). "Comparison of speech perception in background noise with acceptance of background noise in aided and unaided conditions," J. Speech Lang. Hear. Res. **47**, 1001–1011.

Nabelek, A. K., Tucker, F. M., and Letowski, T. R. (**1991**). "Toleration of background noises: Relationship with patterns of hearing aid use by elderly persons," J. Speech Hear. Res. **34**, 679–685.

Onishi, S., and Davis, H. (**1968**). "Effects of duration and rise time of tone bursts on evoked potentials," J. Acoust. Soc. Am. **44**, 582–591.

Petri, A. (**1960**). "Some psychological aspects of pain and the relief of suffering," Ann. N.Y. Acad. Sci. **86**, 13–27.

Petrie, A. (**1967**). *Individuality in Pain and Suffering* (Chicago Press, Chicago).

Picton, T. W., Hillyard, S. A., Krausz, H. I., and Galambos, R. (**1974**). "Human auditory evoked potentials. I. Evaluation of components," Electroencephalogr. Clin. Neurophysiol. **36**, 179–190.

Rajan, R. (**1991**). "Protective function of the efferent pathways to the mammalian cochlea: A review," in: *Noise Induced Hearing Loss*, edited by A. Dancer, D. Henderson, R. Salvi, and R. Hamernik (Mosby, St. Louis), pp. 45–59.

Rogers, D. S., Harkrider, A. W., Burchfield, S. B., and Nabelek, A. K. (**2003**). "The influence of listener's gender on the acceptance of background noise," J. Am. Acad. Audiol **14**, 372–382.

Rueckert, L., and Grafman, J. (**1996**). "Sustained attention deficits in patients with right frontal lesions," Neuropsychologia **34**, 953–963.

Shucard, J. L., Shucard, D. W., Cummins, K. R., and Campos, J. J. (**1981**). "Auditory evoked potentials and sex-related differences in brain development," Brain Lang **13**, 91–102.
Stapells, D. (**2002**). "Cortical event-related potentials to auditory stimuli," in *Handbook of Clinical Audiology*, edited by J. Katz (Lippincot Williams & Wilkins Baltimore, MD), pp. 378–406.
Stern, C. E., Sherman, S. J., Kirchhoff, B. A., and Hasselmo, M. E. (**2001**). "Medial temporal and prefrontal contribution to working memory tasks with novel and familiar stimuli," Hippocampus **11**, 337–346.
Tampas, J. W., Harkrider, A. W., and Nabelek, A. K. (**2004**). "Physiologic correlates of background noise acceptance," J. Acoust. Soc. Am. **115**, 2500–2501.

A glimpsing model of speech perception in noise

Martin Cooke[a)]
Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, United Kingdom

(Received 18 March 2005; revised 19 December 2005; accepted 19 December 2005)

Do listeners process noisy speech by taking advantage of "glimpses"—spectrotemporal regions in which the target signal is least affected by the background? This study used an automatic speech recognition system, adapted for use with partially specified inputs, to identify consonants in noise. Twelve masking conditions were chosen to create a range of glimpse sizes. Several different glimpsing models were employed, differing in the local signal-to-noise ratio (SNR) used for detection, the minimum glimpse size, and the use of information in the masked regions. Recognition results were compared with behavioral data. A quantitative analysis demonstrated that the proportion of the time–frequency plane glimpsed is a good predictor of intelligibility. Recognition scores in each noise condition confirmed that sufficient information exists in glimpses to support consonant identification. Close fits to listeners' performance were obtained at two local SNR thresholds: one at around 8 dB and another in the range −5 to −2 dB. A transmitted information analysis revealed that cues to voicing are degraded more in the model than in human auditory processing. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2166600]

PACS number(s): 43.66.Ba, 43.71.Es, 43.72.Dv, 43.66.Dc [DOS] Pages: 1562–1573

I. INTRODUCTION

Visually, objects are frequently identified based on partial views due to occlusion by other objects. However, the occlusion metaphor is not so obvious in hearing since the contributions from different acoustic objects combine additively in the sound mixture reaching the ears. Consequently, many perceptual and engineering studies have examined the separation of a target signal from a collection of sources that make up the background. Computational attempts at speech separation have been inspired by auditory scene analysis (Bregman, 1990; Cooke and Ellis, 2001), source independence (Comon, 1994; Hyvarinen *et al.*, 2001) and prior source models (Varga and Moore, 1990; Gales and Young, 1993). In practice, none of these approaches has been successful in extracting complex signals such as speech in everyday adverse conditions representative of those faced by listeners, since masking at the auditory periphery complicates the estimation of the energy contribution of the target speech source in each time–frequency region.

Two characteristics of speech signals motivate a different approach to understanding how speech might be recognized in noise. First, since speech is a highly modulated signal in time and frequency, regions of high energy are typically *sparsely distributed*. Consequently, the spectrotemporal distribution of energy in a mixture contains regions that are dominated by the target speech source, even at quite adverse signal-to-noise ratios (SNRs). In such regions, the "noisy" energy observations are very close to those in clean speech, rendering the problem of energy separation unnecessary. Second, the information conveyed by the spectrotemporal energy distribution of clean speech is *redundant*, as demonstrated by numerous studies of the intelligibility of speech after spectral filtering (Fletcher, 1953; Warren *et al.*, 1995; Lippmann, 1996; Kasturi *et al.*, 2002), temporal gating or modulation (Miller and Licklider, 1950; Strange *et al.*, 1983; Gustafsson and Arlinger, 1994), or spectrotemporal impoverishment (Drullman, 1995; Shannon *et al.*, 1995). Redundancy allows speech to be identified based on relatively sparse evidence.

Sparseness and redundancy give rise to an account of speech perception in noise based on the use of "glimpses" of speech in spectrotemporal regions where it is least affected by the background. Figure 1 depicts the regions of a modeled spectrotemporal excitation pattern (STEP; Moore, 2003) dominated by speech in the presence of three different maskers at a global SNR (i.e., SNR measured over the entire token) of −6 dB. Details of the STEP computation are given in Sec. IV B 1. The speech token, shown in the upper panel, is the syllable /ɑrɑ/ spoken by a male. The three panels in the middle row show, from left to right, the token masked by a single talker, eight-talker babble, and speech-shaped noise, respectively. The lower panels depict potential glimpses of the speech target in each of these masking conditions. In this figure, glimpses are defined as those spectrotemporal regions, where the speech energy exceeds that of the masker by at least 3 dB, although the effect of other values of this threshold are reported later in the paper. A substantial proportion of regions are unaffected by the masker, even at an adverse global SNR. Figure 1 also suggests why global SNR is not, on its own, a good predictor of intelligibility. Even though the global SNR is identical in all masking conditions, the typical glimpse size differs. Many studies have demonstrated that intelligibility at a fixed global SNR depends on the type of masker used. A single competing talker or amplitude-modulated noise is a far less effective masker than

[a)]Telephone: +44 114 2221822; Fax: +44 114 2221810. Electronic mail: m.cooke@dcs.shef.ac.uk

 0001-4966/2006/119(3)/1562/12/$22.50

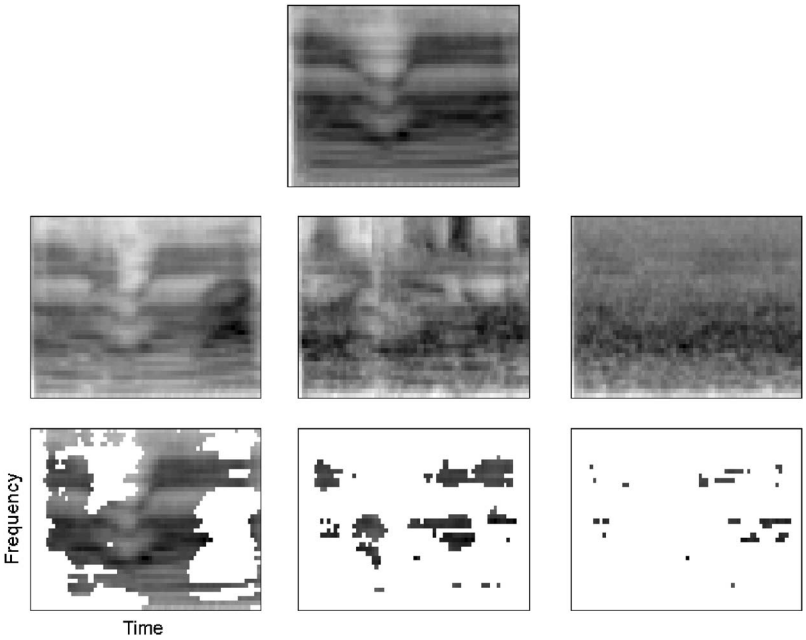


FIG. 1. An illustration of potential glimpses of a short speech token in three masking conditions. Top: spectrotemporal excitation pattern (STEP) for the speech token /ɑrɑ/. Middle: STEP representations for the mixture of the clean speech token with speech from a single competing talker (left), eight talkers (middle), and speech-shaped noise (right). Bottom: regions of the STEP, where the energy of the speech token exceeds that of the masker by at least 3 dB.

multispeaker babble or speech-shaped noise (Miller, 1947; Festen and Plomp, 1990; Scott *et al.*, 2004; Simpson and Cooke, 2005).

In the single competing talker condition, much of the evidence for /ɑrɑ/ survives in large chunks since the masker has significant spectral and temporal energy modulations. Fewer elements remain for the eight-talker masker due to filling in of the energy dips. However, the structures that survive in this example retain evidence of formant movement in the /ɑ/ to /r/ transition as well as intermittent evidence of vowel formants. In contrast, glimpses in the speech-shaped noise background are sparse and small at this SNR due to the absence of significant spectrotemporal modulations in the masker.

The notion that listeners use glimpses to identify speech in noise has been invoked by several researchers (Miller and Licklider, 1950; Howard-Jones and Rosen, 1993; Assmann and Summerfield, 1994; Culling and Darwin, 1994; Assmann, 1996) and has received renewed attention in recent years (Buss *et al.*, 2003; Buss *et al.*, 2004; Cooke, 2003; Assmann and Summerfield, 2004; Alcantara *et al.*, 2004; Freyman *et al.*, 2004).

Miller and Licklider (1950) measured the intelligibility of words in sentences whose waveforms had been interrupted by gating them on and off at a range of modulation frequencies. Across a broad range of interruption rates centered around 15 Hz, word identification scores were almost at the level of uninterrupted speech, in spite of the loss of 50% of the waveform. Miller and Licklider suggested that listeners were able to identify sentences by piecing together glimpses of the target speech available in the uninterrupted portions. In the Miller and Licklider study, glimpses were synchronous in that the complete spectrum was available. Using a "checkerboard" noise masker, Howard-Jones and Rosen (1993) investigated whether listeners could identify consonants in conditions that required the integration of asynchronous glimpses. By varying the spectral extent of each "checkerboard square," Howard-Jones and Rosen demonstrated that listeners were able to integrate glimpses occurring asynchronously, but only for wide spectral regions, accounting for around a quarter to a half of the frequency region of importance for speech perception. Buss *et al.* (2003) studied the effect on spondee identification of amplitude modulation (AM) coherence of either a noise masker or the speech signal filtered into nine narrow bands. They found that the identification of masked AM speech did not depend on the coherence of the modulating waveforms across frequency bands, suggesting that listeners are capable of piecing together relatively narrow spectral glimpses of speech occurring asynchronously. Buss *et al.* (2004) confirmed and extended this finding using a consonant identification task in 16 frequency bands.

The studies of Miller and Licklider (1950), Howard-Jones and Rosen (1993), and Buss *et al.* (2003, 2004) provided a temporal window of glimpsing opportunities whose duration was of the order of a phoneme. Since the most energetic regions of speech occur primarily during voiced episodes, it is reasonable to assume that temporal modulations at the mean voiced–unvoiced alternation rate of speech lead to phoneme-sized intervals of dominance, at least in those spectral regions occupied by formants. Other investigators have proposed that glimpsing opportunities of rather shorter durations may be exploited by listeners. Culling and Darwin (1994) suggested that waveform interactions that give rise to envelope modulation or "beating" provide brief glimpsing opportunities that allow listeners to identify one or other member of a simultaneous vowel pair. They showed that a vowel classifier operated on a sliding 10 ms temporal window at the output of a filterbank analysis could account for listeners' identification rates for vowels whose fundamental frequencies (F0s) were similar enough to produce significant beating. Assmann (1996) extended the Culling and Darwin model to vowels embedded in CVC syllables. Both studies concluded that brief glimpses of the entire spectral profile that occur as the result of waveform interactions can support the identification of vowels with close F0s. However,

the example in Fig. 1 suggests that naturally-produced speech in a range of background sources does not result in many temporal regions where information across the entire frequency spectrum is glimpsed. Instead, brief glimpses of *partial* spectral information do occur.

The intelligibility of speech resynthesized from partial information was measured by Roman *et al.* (2003) in order to assess the performance of an algorithm that used location cues to separate the contributions from two or three sources positioned in a number of spatial configurations. They found that resynthesis from partial spectrotemporal information led to large speech intelligibility improvements over the unprocessed mixture. Similarly, a recent study by Brungart *et al.* (submitted) measured the intelligibility of speech resynthesised from fragments similar to those depicted in Fig. 1. They demonstrated that intelligibility remained at high levels, even for putative glimpses of a speech target in a background composed of four competing talkers. The Roman *et al.* (2003) and Brungart *et al.* (submitted) studies suggest that sufficient information may exist in glimpses to support human speech perception. However, the task of identifying speech synthesized from partial, but clean, information is somewhat different from that faced by listeners, who have the additional problem of identifying which parts of a noisy signal should be treated as glimpses of the target speech.

The current study adopted a complementary approach to that of Brungart *et al.* by comparing the outputs of a computational model of glimpsing with listeners' performance on the same task (Simpson and Cooke, 2005). This approach allowed different models for the detection and integration of glimpses to be evaluated with respect to behavioral data. The glimpsing model employed missing-data algorithms derived from those used in robust automatic speech recognition (Cooke *et al.*, 1994; Cooke *et al.*, 2001), reviewed in Sec. II. Section III describes the speech in noise task and behavioral results, and provides a quantitative analysis of glimpsing opportunities afforded by a number of different maskers. The glimpse detection and identification model is described in Sec. IV.

II. REVIEW OF AUTOMATIC SPEECH RECOGNITION WITH MISSING DATA

Missing-data automatic speech recognition (ASR) was introduced by Cooke *et al.* (1994) as a technique for handling partial data in robust ASR. It has subsequently been applied as a computational model of vowel identification (de Cheveigné and Kawahara, 1999), in the recognition of sine-wave speech (Barker and Cooke, 1997) and as a model of narrow band speech perception (Cunningham, 2003). It has been used as a component of engineering systems for robust ASR to handle additive noise (Drygajlo and El-Maliki, 1998; Raj *et al.*, 1998) and reverberation (Palomaki *et al.*, 2002). Missing-data recognition also forms the core of a probabilistic decoder for multiple acoustic sources (Barker *et al.*, 2005).

Although missing-data techniques can be used in many ASR architectures, in this section it is shown how they can be applied to the most commonly used approach, namely hidden Markov models (HMMs). A HMM is typically used to model units of speech such as phones, triphones, syllables or words, and consists of a number of states, each of which models some segment of the acoustic waveform. States are linked by directed arcs that indicate allowable state transitions, and each transition has a probability associated with it. The relationship between HMM states and the speech waveform is not fixed in advance. Instead, the model learns both the state transition probabilities and probability distributions for each state from a large amount of training data. Each state probability distribution represents an estimate of the process that generated a given segment of the waveform corresponding to that state. Prior to speech recognition using HMMs, the waveform is transformed into a sequence of parameter vectors, each of which corresponds to some short segment of the waveform. Here, parameter vectors represent modeled auditory excitation patterns of the kind shown in Fig. 1.

The core process during recognition is a computation of the likelihood $f(x|C_i)$ that a HMM state C_i could have generated a parameter vector x. This process is repeated for each HMM state and each vector in the sequence, and the HMM that contains the most likely sequence of states is treated as the most likely model for the observed waveform. For the *continuous density* HMMs used in this study, the probability distribution is constrained to be a mixture of Gaussians,

$$f(x|C_1) = \sum_{k=1}^{M} P(k|C_i) f(x|k, C_i), \tag{1}$$

where the $P(k|C_i)$ are the mixture coefficients.

For missing-data ASR using continuous density HMMs, the likelihood computation step is modified as follows. The parameter vector x now consists of components that are regarded as clean and hence reliable, x_r, and other components that are masked, and hence unreliable, x_u. It can be shown (Cooke *et al.*, 2001) that, under a number of assumptions of the kind normally applied in ASR, the required likelihood is given by

$$f(x_r|C_i) = \sum_{k=1}^{M} P(k|C_i) f(x_r|k, C_i) \int f(x_u|k, C_i) dx_u. \tag{2}$$

If the unreliable components are missing altogether (i.e., no information about them is available, as might be the case if the signal had been filtered), then the integral in Eq. (2) reduces to unity, and the likelihood is evaluated as a weighted sum of partial likelihoods,

$$f(x_r|k, C_i) = N(x_r; \mu_{r,k,i}, \sigma^2_{r,k,i}), \tag{3}$$

where $N(x; \mu_{k,i}, \sigma^2_{k,i})$ denotes the Gaussian distribution for mixture component k in state i. Means, variances, and mixture coefficients are estimated during the training process. Equations (2) and (3) forms the basis for the *glimpses-only* model, described in Sec. IV B 3.

There are situations where the unreliable components do contain information. If the parameter vector x represents an energy-based quantity such as the modeled auditory excitation patterns used in this study, and if the unreliability is caused by masking from a noisy background, then the observed value can be treated as an upper bound, x_{high}, on the

energy of the masked speech component at that point in time and frequency. By taking some fixed lower bound for energy, x_{low}, it can be shown (Cooke *et al.*, 2001) that the integral in Eq. (2) is the difference of error functions:

$$\int f(x_u|k,C_i)dx_u = \frac{1}{2}\left[\text{erf}\left(\frac{x_{\text{high},u}-\mu_{u,k,i}}{\sqrt{2}\sigma_{u,k,i}}\right) - \text{erf}\left(\frac{x_{\text{low},u}-\mu_{u,k,i}}{\sqrt{2}\sigma_{u,k,i}}\right)\right]. \quad (4)$$

This computation forms the basis for the *glimpses-plus-background* model described in Sec. IV B 3.

III. SPEECH IN NOISE TASK

A. Speech and noise material

Speech tokens were drawn from the vowel-consonant-vowel (VCV) corpus collected by Shannon *et al.* (1999). The subset of 16 consonants /b, p, d, t, g, k, m, n, l, r, f, v, s, z, ʃ, tʃ/ in the vowel context /ɑ/ was employed. Of the 10 repetitions of each VCV from each talker,[I] 2 from each of 5 male talkers made up a test set of 160 items. The remaining 8 tokens of each VCV were used as a training set for the recognizer. Tokens were normalized to have equal rms energy.

Noise signals were formed by modulating speech-shaped noise with the envelope of *N*-talker babble for various *N*. Such signals produce approximately the same amount of energetic masking as the signal from which the envelope is drawn (Festen and Plomp, 1990) and were used here to limit informational masking effects. Following Brungart *et al.* (2001), the envelope was computed by convolving the absolute value of an *N*-talker babble signal with a 7.2 ms rectangular window. Babble was produced by summing utterances with equal rms energy from the TIMIT corpus (Garofolo *et al.*, 1992). Twelve babble-modulated noise conditions were employed corresponding to the following values of *N*: 1, 2, 3, 4, 6, 8, 16, 32, 64, 128, 512, and ∞ (i.e., speech-shaped noise).

Noisy tokens were produced by adding babble-modulated speech-shaped noise in each of the 12 noise conditions to the test part of the corpus at a global SNR of −6 dB, computed on a token-by-token basis. Since the energetic content of VCV syllables is dominated by the vowel portions, the actual SNR in the consonant portions was somewhat lower than −6 dB. Noise waveforms were gated on and off with the speech tokens.

B. Quantitative analysis of glimpsing opportunities

A quantitative analysis of glimpsing opportunities as a function of *N* was conducted. For this analysis, a glimpse was defined as a connected region of the spectrotemporal excitation pattern in which the energy of the speech token exceeded that of the background by at least 3 dB in each time–frequency element or "pixel." A "pixel" corresponds to a single time frame and frequency channel in the STEP representation. Here, elements of a region were deemed to be connected if they were part of the four-neighborhood (i.e., excluding diagonal neighbors) of any other element in the region. The "area" or extent of a glimpse was taken to be the number of time–frequency elements making up the glimpsed region. This is not an area in the traditional sense since the time and frequency units are not identical. Further, different choices of time and frequency resolution in the STEP will result in slight differences in calculated "areas." The choices here (defined in Sec. IV B 1) are based on those commonly used in studies employing STEPs.

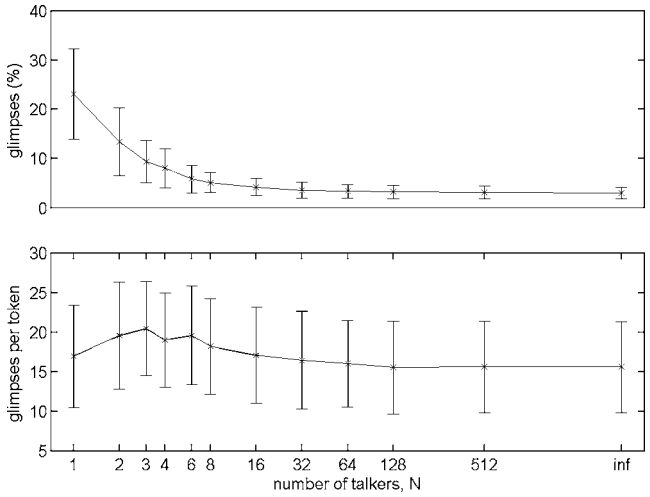


FIG. 2. Glimpse percentages and counts expressed as means across all tokens in the test set, separately computed for each noise condition. Error bars represent ±1 standard deviation. Here and elsewhere, the point marked "inf" indicates the speech-shaped noise condition.

Two quantities—glimpse area and glimpse count—were measured for each noise token in the corpus. Means and standard deviations of the two measures in each noise condition are shown in Fig. 2. The upper panel displays the mean percentage of each token covered by glimpses. While there is substantial variability across individual tokens, the mean glimpse percentage falls rapidly with *N*, leveling off at around 3% for $N>16$. The lower panel of Fig. 2 plots the number of glimpses per token. Interestingly, the range of means is quite small, suggesting that each noise condition results in a similar number of glimpses, although the quality of opportunities (as defined by the glimpse area) differs substantially as a function of *N*. These metrics confirm that qualitatively different glimpsing opportunities are available as the number of talkers contributing to the babble-modulated masker is varied, as demonstrated in Fig. 1.

C. Behavioral experiment

The intelligibility of consonants in babble-modulated noise was measured by Simpson and Cooke (2005) as part of a study comparing the masking effectiveness of natural babble and babble-modulated noise for varying numbers of talkers contributing to the babble. The babble-modulated noise conditions and speech tokens used by Simpson and Cooke (2005) were precisely the same as those used in the current study, as described in Sec. III A. Consonant identification scores based on 12 normal-hearing listeners are replotted in Fig. 3. In line with an earlier study that employed babble-modulated noise (Bronkhorst and Plomp, 1992), in-

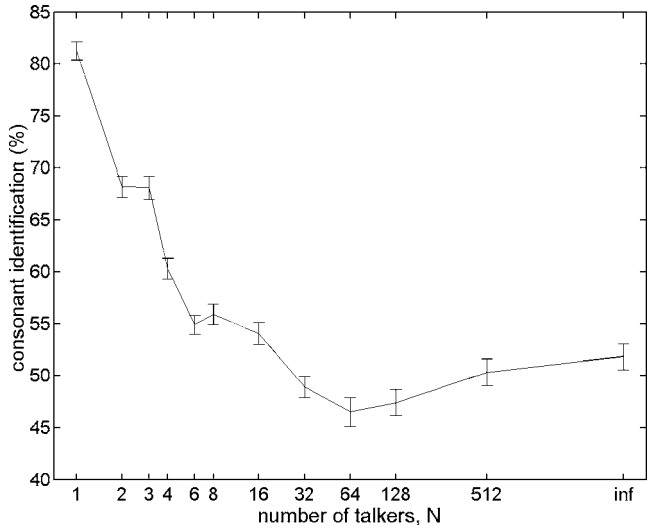


FIG. 3. Consonant intelligibility as a function of the number of talkers in the babble-modulated noise. Error bars represent ±1 standard error. Replotted from Fig. 1 of Simpson and Cooke (2005).

telligibility falls steeply with the number, N, of contributing talkers before leveling out for $N>32$. In fact, the intelligibility for all $N>6$ is not significantly different from speech-shaped noise.

Figure 4 plots the mean glimpse percentage in each noise condition against listeners' intelligibility results. The high correlation (0.955) between these measures suggests that the glimpse proportion alone is a very good linear predictor of intelligibility for these stimuli.

IV. GLIMPSING MODEL

A. Overview

Figure 5(a) illustrates the architecture of the glimpsing model for listeners. Glimpse detection is assumed to operate on a spectrotemporal representation produced in the early stages of auditory processing. Glimpses then have to be tracked through time and integrated to form a running speech

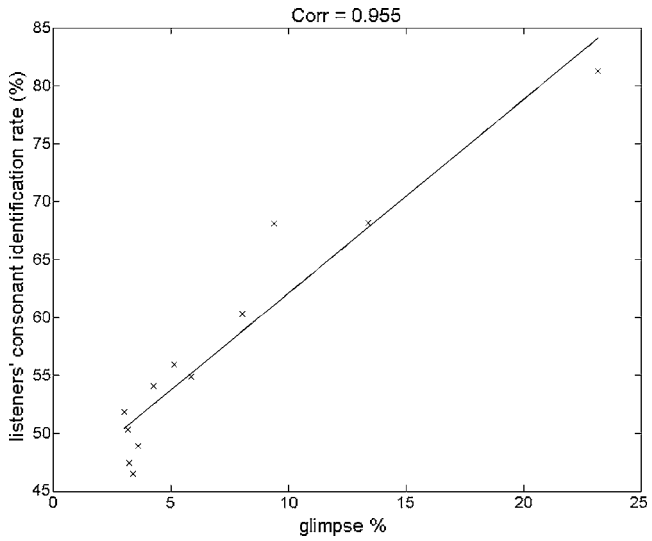


FIG. 4. The correlation between intelligibility and proportion of the target speech in which the local SNR exceeds 3 dB. Each point represents a noise condition, and proportions are means across all tokens in the test set. The best linear fit is also shown. The correlation between listeners and these putative glimpses is 0.955.

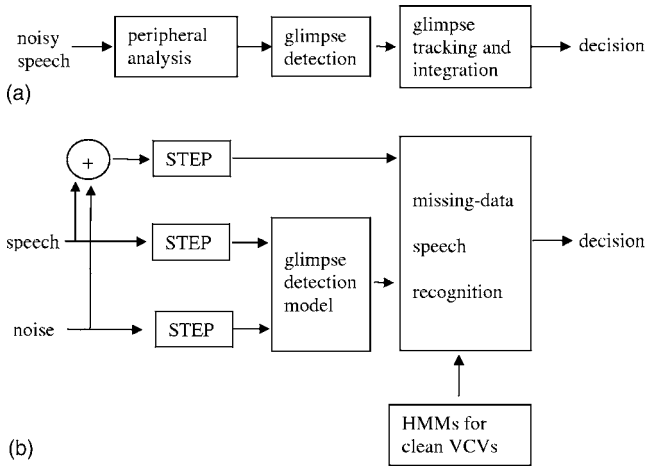


FIG. 5. (a) Architecture of the glimpsing model in humans. (b) Processing steps in the current study.

hypothesis. The temporal integration of glimpses may be based on estimates of pitch movements or speaker characteristics such as vocal tract length. Since the purpose of this study was to explore glimpse detection models that best account for behavioral findings rather than to provide a computational procedure for robust automatic speech recognition in noise, the actual processing steps used to test the model had access to both speech and noise signals [Fig. 5(b)]. STEPs for speech, noise, and their mixture were computed independently. The glimpse detection model determines those spectrotemporal regions dominated by speech. These regions constitute a mask that, along with the STEP for the mixture, was processed by a missing-data speech recognizer that computes the most likely speech hypothesis based on the partial evidence provided by the glimpses. The recognizer employed continuous density hidden Markov models derived by training on clean VCV tokens.

B. Methods

1. STEP computation

The spectrotemporal excitation pattern used as input to the model is a smoothed and compressed representation of the envelope of the basilar membrane response to sound. The waveform was initially processed by a bank of 40 gammatone filters (Patterson *et al.*, 1988) implemented using a pole-mapping procedure described in Cooke (1993). Filters were equally spaced on an ERB-rate scale between 50 and 7500 Hz. The Hilbert envelope in each channel of the filterbank was computed and smoothed with a leaky integrator with an 8 ms time constant (Moore *et al.*, 1988). The smoothed envelope was downsampled to 100 Hz and log-compressed.

2. Glimpse detection

One of the purposes of the current study was to explore the effect of different assumptions about what constitutes a usable and detectable glimpse. The glimpses shown in Fig. 1 result from a detection model which assumes that all spectrotemporal elements whose local SNR exceeds 3 dB are detectable and used in a subsequent classification of the pat-

tern. However, there is no prior reason for choosing this particular local SNR threshold. Further, the assumption that listeners can detect very small regions of favorable local SNR when surrounded by a masker may be unreasonable. The recognition experiments reported here treated both local SNR for the detection and minimum glimpse area as free parameters over which the match to listeners' performance was optimized. The output of the glimpse detection stage is a time–frequency mask.

3. Missing-data recognition

The glimpse mask and STEP representation of the mixture signal provide the input to a missing-data speech recognizer. This study employed the HMM-based missing-data methods described in detail by Cooke *et al.* (2001) and reviewed in Sec. II. Two variants of the missing-data technique were used. The first, referred to as the glimpses-only model, employed information in the glimpsed regions alone and ignored information in the masked regions. The second, the glimpses-plus-background model, used, in addition, information from the masked regions. The glimpses-plus-background model was motivated by the phoneme restoration effect (Warren, 1970) and the subsequent demonstration of spectral induction (Warren *et al.*, 1997). These studies used speech with missing temporal fragments or spectral regions, respectively. In both cases, listeners were more likely to perceive or identify sounds if the missing parts were replaced by noise with sufficient energy to have masked the missing portion, had it been present.

The missing-data recognizer used a continuous density HMM for each of the 16 VCVs. Each VCV model had six emitting states, while each state employed a four-component Gaussian mixture density. Here, 40 examples of each VCV (8 from each of the 5 male talkers, distinct from those used for evaluation) were used to train the HMMs. Model training was performed using the HMM toolkit HTK (Young and Woodland, 1993) while recognition used a local implementation of a missing-data recognizer. Performance in the no-noise condition exceeded 99%. For the glimpses-plus-background model, a lower energy bound of −120 dB was used for x_{low} in Eq. (4).

C. Results

1. Variation in local SNR threshold for glimpse detection

Figure 6 compares listeners' and model recognition rates in each noise-masking condition, expressed as the percentage of consonants correctly identified across the test set, for a single local SNR detection threshold. The closeness of the match obtained for this particular choice of detection model suggests that glimpsing can account for the gross pattern of intelligibility of VCVs in stationary and nonstationary backgrounds.

Figure 7 compares listeners' and model recognition rates as a function of the local SNR for glimpse detection. The upper panel plots the root-mean-square (rms) listener–model distance computed across noise-masking conditions (e.g., the rms distance between the two curves shown in Fig. 6). This distance is a positive quantity and measures the proximity of

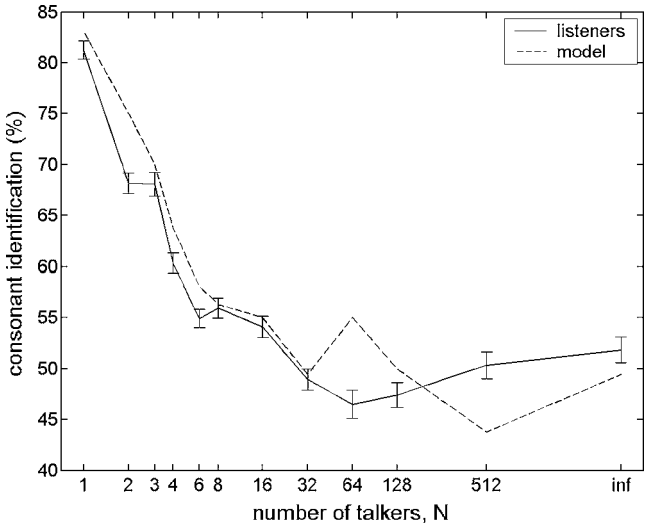


FIG. 6. Model and listeners' consonant identification rates for the glimpses-only model at a local SNR detection threshold of 6 dB. Listeners' data here and in Fig. 9 are the same as that of Fig. 3.

the two curves. The lower panel plots the mean listener-model difference or "bias." This is a signed quantity that reflects the extent to which the model outperforms listeners (positive bias values) or *vice versa* (negative bias values).

Both the glimpses-only and glimpse-plus-background models result in a pair of local minima in the model-listener rms distance. A consideration of the effect of changing the local SNR for detection explains the presence of two minima. At increasingly large negative local SNRs, many regions are treated as glimpses. However, some of these will be dominated by the background source. Consequently, model performance is dominated by errors due to distortions in the glimpsed data and its performance falls below that of listeners, as shown by the bias curve in the lower panel of Fig. 7. At the other extreme of large positive local SNRs for detection, glimpse scarcity limits performance since the number of regions satisfying increasingly strict criteria for inclusion falls as the local SNR increases. Again, the bias curve indicates that model performance suffers relative to listeners in such conditions.

Between the two extremes of glimpse corruption and glimpse scarcity, there is a broad region (from −2 to +6 dB for the glimpses-only model and from −5 to +6 dB for the glimpses-plus-background model), where the model outperforms listeners. Local minima in the listener–model distance occur at the edges of the range. This result suggests that there is more than sufficient information in glimpses to support their use as the basis for speech perception in noise, but that listeners do not possess an optimal glimpse detection model.

The broad pattern of results is similar for both glimpses-only and glimpses-plus-background recognition variants. However, the curves differ in several details. The glimpses-plus-background model outperforms the glimpses-only model, as shown by the more extensive range of SNRs with positive bias in Fig. 7, reflecting the additional information used in the classification process. The closest listener-model match at negative SNR thresholds are of a similar size in the two models, but occur at different points (−2 vs −5 dB), perhaps because the additional information employed by the

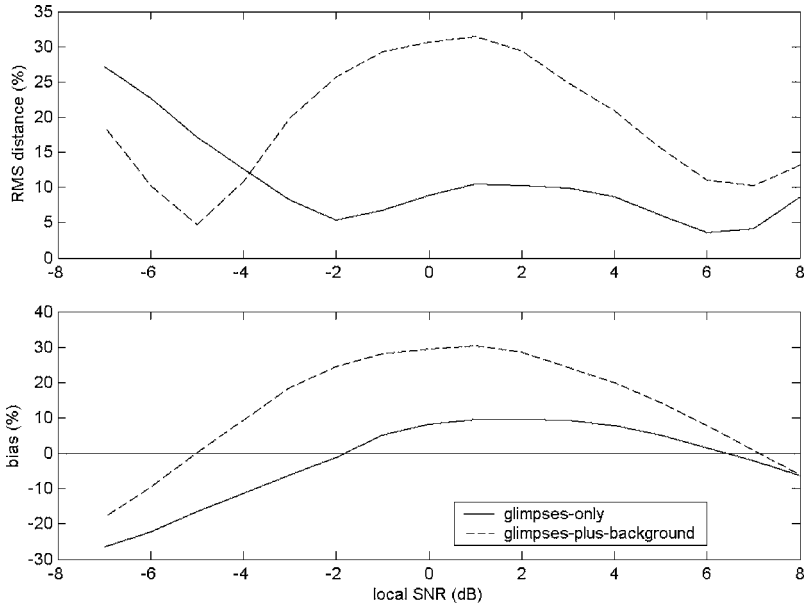


FIG. 7. The rms distance (upper panel) and mean difference (lower panel) between the listeners and model performance across all noise conditions as a function of the local SNR threshold for glimpse detection for the glimpses-only and glimpses-plus-background models.

glimpses-plus-background model allows it to tolerate more distortion in the evidence considered to be part of the speech hypothesis.

2. Variation in glimpse area for detection

Since the basic glimpse detection model outperforms listeners across a broad range of local SNRs for detection, a number of more sophisticated detection models that further reduce the set of glimpses could also account for the results. Based on the assumption that listeners may be unable to detect very brief regions of target dominance, or regions that occupy a very narrow portion of the spectrum, a minimum glimpse area criterion was incorporated into the glimpse detection model. Specifically, all connected regions of spectrotemporal elements satisfying a given local SNR criterion also had to possess an "area" (as defined in Sec. III B) greater than a specified amount.

Figure 8 plots the rms model-listener distance as a function of the local SNR threshold and minimum glimpse area. An inspection of the bias functions illustrates that when glimpses with a minimum area of 25 were removed, the glimpses-only model underperformed listeners, while the glimpse-plus-background variant tolerated removal of glimpses with 75–100 elements before performance fell be-

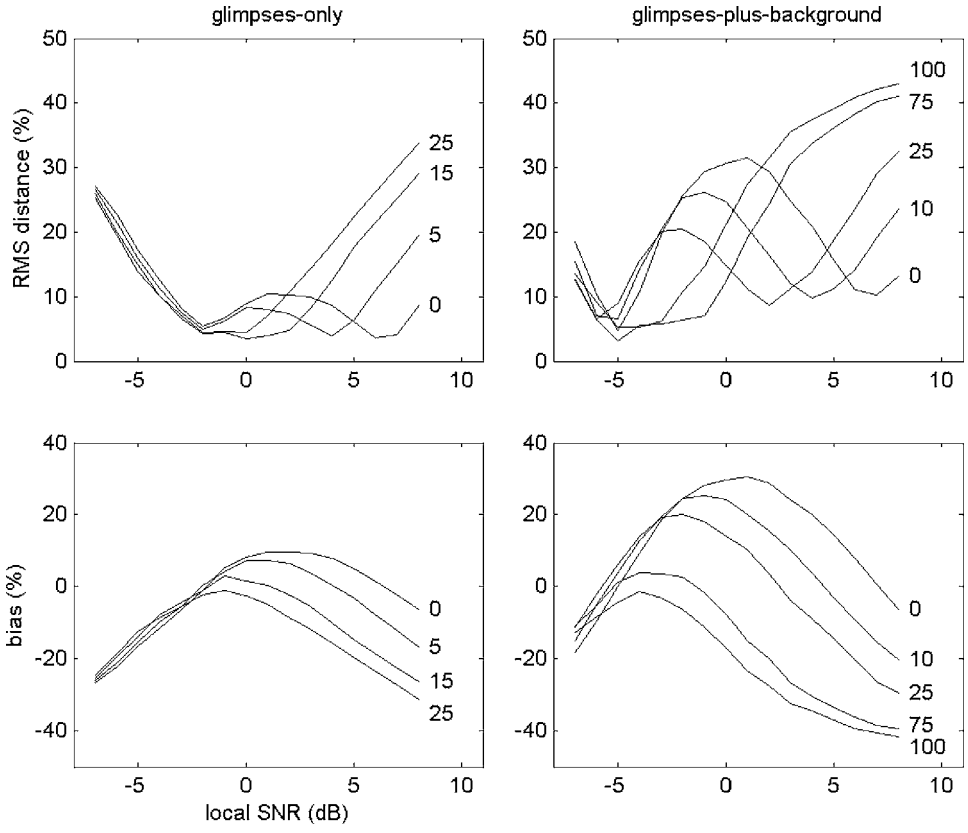


FIG. 8. The rms distance and bias for glimpse detection models in which both the local SNR for detection and the minimum glimpse area are varied. Numbers attached to curves identify minimum glimpse areas used.

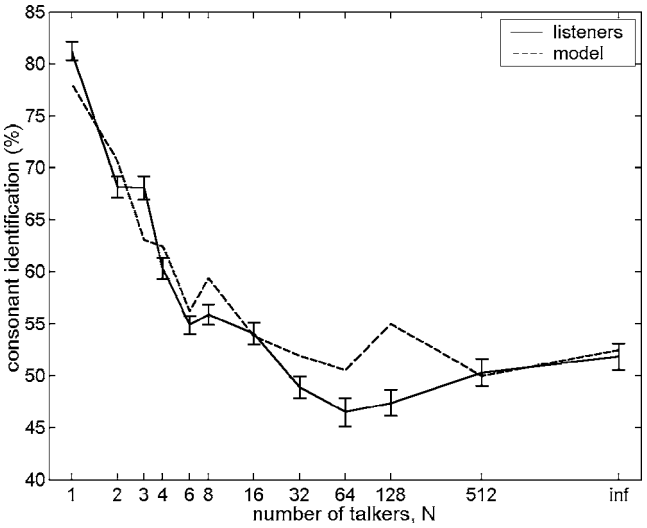


FIG. 9. Best listener-model fit (glimpses+background variant, local SNR for detection=−5 dB, minimum glimpse areas=75).

low that of listeners. The superior robustness of the latter model demonstrates that masked regions contain useful information for the discrimination between hypotheses about the identity of the speech target, in spite of the fact that no energy separation is performed in such regions. A striking feature of the results is the sensitivity of the best-fit location at positive local SNRs. As increasingly larger glimpses are excluded, the position of the minimum rapidly shifts to lower local SNRs, presumably because of the decreasing amount of evidence on which to base identification. In contrast, the location of the other local minimum barely changes as data are excluded. This finding suggests that the best fit at a negative glimpse detection threshold is robust, even though some glimpses are corrupted by energy from the background. The best fit occurred at a local SNR for detection of −5 dB within a minimum glimpse area of 75 elements for the glimpses-plus-background model (Fig. 9). All further analyses were based on this best-fitting model.

3. Consonant identification rates and confusions

Figure 10 compares listeners' and model identification rates for each consonant, averaged across noise conditions. While the overall identification rates are very similar (listeners: 57.7%, model: 57.5%), there are clear differences in scores for individual consonants. The model always outperformed listeners in the identification of plosives and nonsibilant fricatives (/v/ and /f/). The difference in performance was sometimes large (e.g., for /b/ and /f/). Listeners scored more highly than the model for the sibilants /s/, /ʃ/, and /tʃ/ and for the nasals. In the case of /m/, the margin was large. The source of these disparities is not clear.

Figure 11 depicts model and listener identification rates for each consonant as a function of noise condition. These results show that the overall fall in intelligibility as *N* increases is not always reflected in the recognition rates of individual consonants. Recognition of the plosives (/t/ and /g/) and the affricate /tʃ/ was uniformly good across differing masking conditions, while /b/ and /v/ were difficult to identify in the presence of most maskers. The distribution of available cues in a noise modulated by the envelope of a single talker will be quite different from those that remain after masking by stationary noise, and it may be that the multiple, redundant cues to identity are better distributed in some consonants than in others.

Tables I and II report consonant confusions for listeners and the best-fitting model, respectively. At the detailed level of individual confusions, it is clear that the model differs from listeners. However, some common trends can be seen. Both model and listeners tend to over-report /t/. Similarly, /m/ and /v/ tokens result in a wide range of response alternatives. Table III shows the proportion of transmitted information for voicing, manner and place for listeners and the best-fitting model (Miller and Nicely, 1955; Wang and Bilger, 1973). This analysis reveals that while manner and place information is transmitted at a similar rate in the listeners and model, voicing information is far more susceptible to masking in the latter. This may be a defect of the excitation pattern representation employed by the model. Implicit cues to voicing are conveyed by the resolved harmonics in the lower-frequency region of the excitation pattern, but temporal cues that would indicate the presence of voicing are lost. Spectral averaging in the formation of HMM probability densities may also contribute to the degradation of voicing information (Barker, 1998).

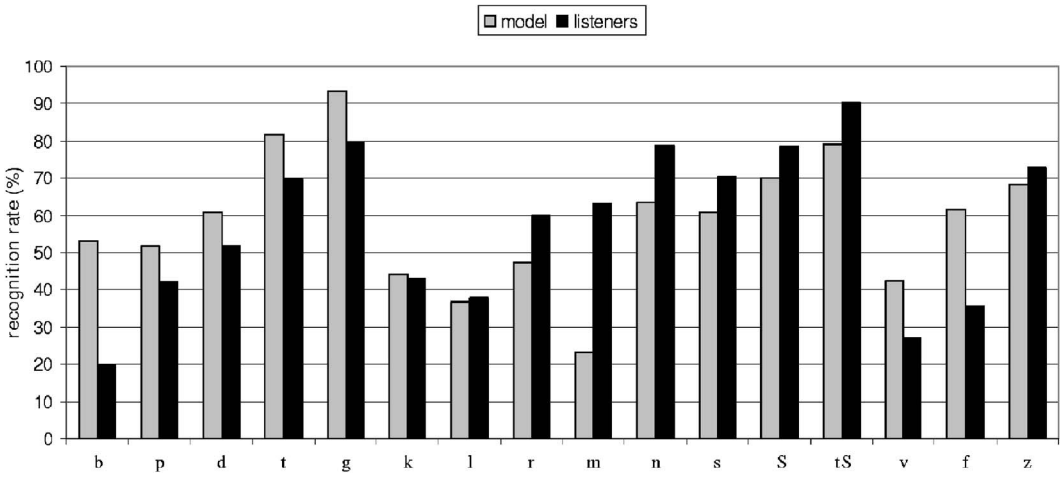


FIG. 10. Consonant identification scores for listeners and the best-fitting model.

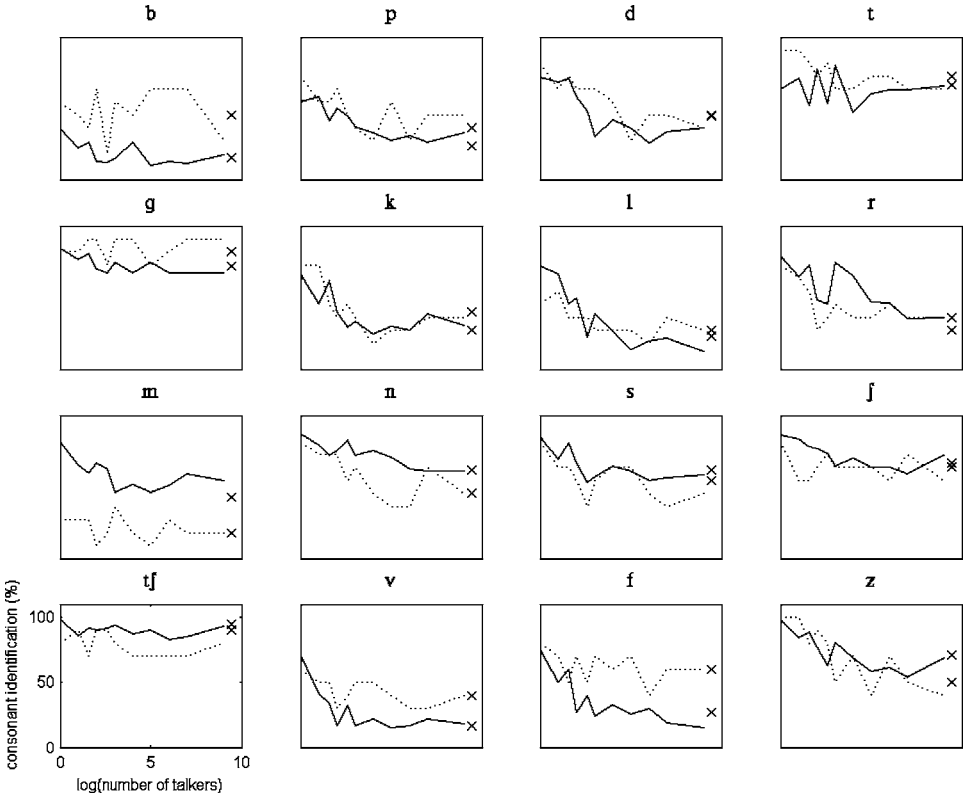


FIG. 11. Consonant identification scores as a function of N (solid line: listeners, dotted line: model). Crosses depict scores in the speech-shaped noise condition.

4. Response set probability

A valid model of speech perception should be able to predict the most likely judgements made by a set of listeners to each individual noisy token in an evaluation set. Such an evaluation criterion is extremely demanding, given our current understanding of speech perception in noise. A suitable metric was developed and evaluated on the current task. The metric computes the log probability of the entire response sequence. The response probability for an individual token can be computed from the discrete probability distribution formed from the set of listener responses to that token. This method requires a sufficiently large number of judges in order to provide a robust estimate of the response probability distribution. The probability of the entire sequence—the response set probability—is the product of the individual response probabilities.

Response set probabilities for each listener and the model can be compared. The probability for any individual listener is obtained by leaving out their judgements and using the judgements of the other listeners to form the probability distribution for each token. Statistical hypothesis testing can

TABLE I. Confusion matrix for listeners, averaged over noise conditions. Numbers represent percentage responses. For clarity, response percentages smaller than 1.5% have been removed. Consequently, column totals may not match the sum of column entries. Confusions greater than or equal to 10% are presented in bold. Each (row, column) entry signifies the (stimulus, response) confusion.

	b	p	d	t	g	k	l	r	m	n	s	ʃ	tʃ	v	f	z
b	**21**	9	5	3	4	4	6	5	7	3				**21**	9	2
p	3	**43**		7		**12**	2	2	7	2				6	**12**	
d			**52**	4	**18**	2	3	2		**11**			2	2		3
t		6	4	**70**		9			2	3					2	
g	2		3		**80**	3		3						2		
k		**14**	2	**10**	6	**43**			4	2			2	2	**10**	
l	5	5	2	2	3		**37**	8	**20**	2				**9**	4	
r	4	4		2	3		5	**59**	9					7	4	
m	3	4		2	2	2	7	4	**63**	7				3		
n			5	3	3		3		4	**79**						
s		2		2		2					**71**				9	9
ʃ												**79**	**15**			
tʃ				2								4	**90**			
v	**12**	7	3	3	4	3	7	7	**14**	2				**27**	6	4
f	6	**23**		4	2	9		2	8					8	**35**	
z			8	3	2	2	2			2			2	2		**73**

TABLE II. Confusion matrix for the best-fitting model.

	b	p	d	t	g	k	l	r	m	n	s	ʃ	tʃ	v	f	z
b	**53**		**12**			**11**	3						6	8	5	
p	3	**52**		**17**			6	4					8	8		3
d	5		**61**	9	8	4							8	3		
t		8	3	**82**		3							3			2
g			3	3	**93**											
k	9	6	3	3	3	**44**		4			3		**13**	3	8	
l	**10**		3	7		**21**	**37**	**17**						3	2	
r	6	3	6	4	2	2	**10**	**48**			**11**			4	2	3
m		**13**	2	**18**		2	4	6	**23**	7	3	3	6	7		8
n				**16**	**10**	2			2	**63**				4		2
s				**22**							**61**					**18**
ʃ												**70**	**27**	2		
tʃ												**18**	**79**			
v	3	6	4	8		9	**10**				4		9	**43**	2	2
f	4		8	5							5	2		**13**	**62**	
z	5	3		7									**10**	5		**68**

TABLE III. Proportion of transmitted information for voicing, manner, and place. Consonant subsets used for each distinction are indicated.

	Listeners	Model
Voicing [b d g l r m n v z] [p t k s ʃ tʃ f]	0.424	0.292
Manner [b d g p t k] [tʃ] [l r] [m n] [v z s ʃ f]	0.397	0.351
Place [b p m f v] [d t n s l z] [g k] [r ʃ tʃ]	0.371	0.367

then be used to determine the probability that the model belongs to the listeners' distribution of response likelihoods.

Since the current study used a relatively small number of listeners, it is difficult to obtain robust estimates of per token response probabilities. One specific issue is the problem of zero probabilities for certain response categories. For instance, in response to an /s/ token in noise, it may be that no listener in the set responded with /g/, leading to a zero probability for that response. If the model were to respond with /g/, the probability of the entire response set becomes zero. One solution is to replace zero probabilities with a certain minimum value and renormalize the distribution.

Response set probabilities for each noise condition were computed for all listeners and for the best-fitting model. A probability floor of 0.1 was used. Log response set probabilities are plotted in Fig. 12. They show that, in all masking conditions, listeners act as a cohort and are more similar to each other than to the model. Listeners' responses are most similar to each other in masking conditions, which lead to a high overall performance due to the lower incidence of confusions.

V. GENERAL DISCUSSION

The main finding of this study is that glimpses contain more than enough information to support speech recognition in a computational model. This suggests that a glimpsing process may serve as a basis for human speech perception in noise. Several glimpsing models can account for listeners' performance on a consonant in noise task. A model that used solely the information in the putative glimpse regions produced a close fit to the behavioral data by assuming that (i) glimpses are detected if the local SNR is greater than 0 dB, and (ii) regions of positive local SNR that do not possess a sufficient spectrotemporal extent will not be detected. A model that additionally exploited information in the masked regions also matched listeners', data but with quantitatively different assumptions. A robust feature of the latter model was the requirement that glimpses be detectable when the local SNR is greater than −5 dB.

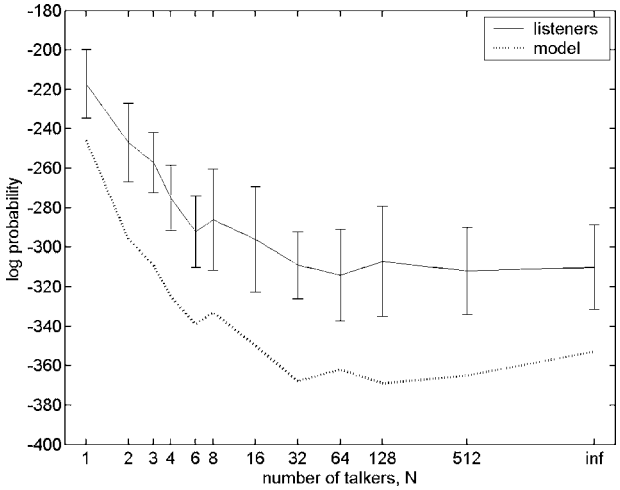


FIG. 12. Log probability of the response set as a function of *N*. Listeners: solid line, model: dotted line. Error bars denote ±2 standard deviations.

The optimum local SNR detection thresholds for the glimpses-only and glimpses-plus-background models suggest somewhat different solutions to the subsequent identification problem. A 0 dB threshold produces fewer glimpses, but those that are available are reasonably undistorted. At a detection threshold of −5 dB, many more regions will be considered as glimpses, but some of them will be significantly contaminated by background energy. The latter value is consistent with the model of Moore *et al.* (1997), which predicts a threshold of −4 dB for detecting complex signals in noise, and with the findings of Brungart *et al.* (submitted), who suggested an optimum local SNR threshold of −6 dB. However, the study of Drullman (1995) found that speech components more than 2 dB below the noise level in a given auditory filter made no contribution to intelligibility.

The glimpse detection models in the current study used simulated spectrotemporal excitation patterns that are good first-order representations of auditory stimuli at an early stage of processing. However, the excitation pattern representations used here are deficient in two respects. They lack both a representation of the fine structure of auditory-nerve fiber response to sound and a model of nonsimultaneous masking. It seems likely that a more realistic model of the temporal fine structure would result in glimpses differing in detail from those in the model presented here. However, the changing local dominance relations between the target and background that give rise to glimpsing opportunities are likely to affect other representations in broadly similar ways. For instance, the temporal response characteristic seen in auditory-nerve fibers discharge patterns to locally strong stimulus components known as "synchrony capture" can be considered as a reflection of the stimulus dominance at the level of a temporal fine structure (e.g., Sinex and Geisler, 1983).

If speech perception in noise is based on glimpsing, listeners must have developed solutions to two fundamental problems—detection and integration. The detection models described here make use of foreknowledge of both local SNR and glimpse extent. Several computational attempts to estimate local SNR have been reported. Berthommier and Glotin (1999) showed that the disruption to harmonicity was a good predictor of local SNR, while Tchorz and Kollmeier (2002) estimated local SNR from amplitude modulation spectrograms using a neural network learning procedure.

Rather than estimating local SNR directly, it is possible that listeners apply principles of auditory organization to group spectrotemporal regions based on properties such as the similarity of location or fundamental frequency (Bregman, 1990; Darwin and Carlyon, 1995; Cooke and Ellis, 2001; Brown and Wang, 2005). An algorithm that could be used for glimpse estimation based on binaural cues was pre-

sented in Roman *et al.* (2003). Alternatively, it might be possible to identify clean fragments using prior speech models.

The second problem is to integrate glimpses into an ongoing speech hypothesis. Assmann and Summerfield (2004) call this the tracking problem. While integration may not represent a significant challenge for a single speech target in modulated babble noise since there is little possibility of confusion between this type of noise and speech, it becomes more important when the background contains other talkers. In such conditions, listeners have to contend with both energetic and informational masking (Carhart *et al.*, 1969; Brungart *et al.*, 2001; Freyman *et al.*, 2004). In the glimpsing account, informational masking can be understood as arising from the incorrect assignment of glimpses to the developing speech hypothesis. Glimpse integration may be less of a problem in practice than detection. A recent computational approach to glimpse integration (Barker *et al.*, 2005) uses prior speech models to track fragments of evidence through time.

The approach adopted here belongs to a small class of models of human speech recognition that exploit speech models and automatic speech recognition techniques (Ainsworth and Meyer, 1994; Ghitza, 1993; Holube and Kollmeier, 1996). ASR provides a well-founded basis for constructing models of speech schema, although it requires the use of larger corpora than are typically employed in perceptual studies. Direct computational models contrast with macroscopic models such as the articulation index (Kryter, 1962), speech transmission index (Steeneken and Houtgast, 1979), and speech intelligibility index (ANSI, 1997), which attempt to predict speech intelligibility in conditions of background noise, reverberation, and channel distortion. In principle, direct computational models can also be applied to speech intelligibility assessment, and promise to provide more detailed information in a wider range of conditions than those handled by macroscopic approaches. For instance, a direct computational model employing a sophisticated statistical model of speech can produce decisions for individual tokens in arbitrarily complex noise backgrounds. However, representations of speech and associated decoding processes used in ASR are likely to be rather different from those employed by listeners. It is clear that current ASR-based models of human speech perception are incapable of predicting listeners' responses in detail. Here, a response set probability metric indicated that the panel of listeners were more similar to each other than to the model, in spite of the existence of a close model-listener match at the level of overall performance. At the level of consonant confusions, clear differences between the current model and behavioral data remain. Information theoretic analyses suggested that the model is weak on the transmission of voicing cues. Whether this represents a deficiency in the spectrotemporal excitation pattern representation itself or stems from a loss of information during the formation of speech schema from individual training tokens is a matter for further study.

VI. CONCLUSIONS

A processing model that takes advantages of relatively clear views of a target speech signal can account for the overall intelligibility of VCV tokens in a range of energetic masking conditions. As a strategy for dealing with speech in noise for listeners and algorithms, focusing on the regions with advantageous local SNR may be simpler than estimating the energy contribution of the speech signal at each point in time and frequency. The best fit to behavioral data came from a model that (i) used information in the glimpses and counterevidence in the masked regions; (ii) restricted glimpses to have a certain minimum area; and (iii) treated all regions with local SNR in excess of −5 dB as potential glimpses. However, many different glimpse detection strategies have the capacity to explain the observed level of listeners' performance.

Attempts to use techniques from automatic speech recognition to model human speech perception are in their infancy, but this study suggests that insights into both ASR and HSR can be gained by a modeling approach that tests different potential perceptual strategies for handling speech in noise.

ACKNOWLEDGMENTS

This work was supported by Grant No. GR/R47400/01 from the UK Engineering and Physical Science Research Council. The author would like to thank Maria Luisa Garcia Lecumberri, Jon Barker, and Sarah Simpson for valuable comments on the manuscript, and Professor Bob Shannon for making available the extended version of the VCV corpus. Three anonymous reviewers provided helpful comments.

[1]The published corpus released a single example of each VCV from every talker, but ten were recorded. Professor Shannon kindly made the original recordings available to the author.

Ainsworth, W. A., and Meyer, G. F. (**1994**). "Recognition of plosive syllables in noise: Comparison of an auditory model with human performance," J. Acoust. Soc. Am. **96**, 687–694.

Alcantara, J. I., Weisblatt, E. J. L., Moore, B. C. J., and Bolton, P. F. (**2004**). "Speech-in-noise perception in high-functioning individuals with autism or Asperger's syndrome," J. Child Psychol. Psychiatry **45**, 1107–1114.

ANSI (1997). American National Standard Methods for Calculation of the Speech Intelligibility Index Document Number: ANSI/ASA S3.5-1997.

Assmann, P. F. (**1996**). "Tracking and glimpsing speech in noise: role of fundamental frequency," J. Acoust. Soc. Am. **100**, 2680.

Assmann, P. F., and Summerfield, Q. (**1994**). "The contribution of waveform interactions to the perception of concurrent vowels," J. Acoust. Soc. Am. **95**, 471–484.

Assmann, P. F., and Summerfield, Q. (**2004**). "The perception of speech under adverse acoustic conditions," in *Speech Processing in the Auditory System*, Springer Handbook of Auditory Research, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (Springer-Verlag, Berlin), Vol. 18.

Barker, J. P. (**1998**). "The relationship between auditory organisation and speech perception: Studies with spectrally reduced speech," unpublished Ph.D. thesis, University of Sheffield.

Barker, J., and Cooke, M. P. (**1997**). "Modelling the recognition of spectrally reduced speech," *Proc. EUROSPEECH*, pp. 2127–2130.

Barker, J., Cooke, M. P., and Ellis, D. P. W. (**2005**). "Decoding speech in the presence of other sources," Speech Commun. **45**, 5–25.

Berthommier, F., and Glotin, H. (**1999**). "A new SNR feature mapping for robust multistream speech recognition," *Proc. XIVth Int. Cong. Phonetic Sciences*, pp. 711–715.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT Press, Cambridge, MA).
Bronkhorst, A. W., and Plomp, R. (**1992**). "Effect of multiple speech-like maskers on binaural speech recognition in normal and impaired hearing," J. Acoust. Soc. Am. **92**, 3132–3139.
Brown, G. J., and Wang, D. (**2005**). "Separation of speech by computational auditory scene analysis, in *Speech Enhancement*, edited by J. Benesty, S. Makino and J. Chen (Springer-Verlag, New York), pp. 371–402.
Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **100**, 2527–2538.
Brungart, D. S., Chang, P., Simpson, B. D., and Wang, D. (submitted). "Isolating the energetic component of speech-on-speech masking with an ideal binary time–frequency mask," J. Acoust. Soc. Am..
Buss, E., Hall, J. W., and Grose, J. H. (**2003**). "Effect of amplitude modulation coherence for masked speech signals filtered into narrow bands," J. Acoust. Soc. Am. **113**, 462–467.
Buss, E., Hall, J. W., and Grose, J. H. (**2004**). "Spectral integration of synchronous and asynchronous cues to consonant identification," J. Acoust. Soc. Am. **115**, 2278–2285.
Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969**). "Perceptual masking in multiple sound backgrounds," J. Acoust. Soc. Am. **45**, 694–703.
Comon, P. (**1994**). "Independent component analysis. A new concept?," Signal Process. **36**, 287–314.
Cooke, M. P. (**1993**). *Modelling Auditory Processing and Organisation* (Cambridge University Press, Cambridge).
Cooke, M. P. (**2003**). "Glimpsing speech," J. Phonetics **31**, 579–584.
Cooke, M. P., and Ellis, D. P. W. (**2001**). "The auditory organization of speech and other sources in listeners and computational models," Speech Commun. **35**, 141–177.
Cooke, M. P., Green, P. D., and Crawford, M. D. (**1994**). "Handling missing data in speech recognition," *Proc. 3rd Int. Conf. Spoken Language Processing*, pp. 1555–1558.
Cooke, M. P., Green, P. D., Josifovski, L., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and uncertain acoustic data," Speech Commun. **34**, 267–285.
Culling, J. F., and Darwin, C. J. (**1994**). "Perceptual and computational separation of simultaneous vowels: cues arising from low frequency beating," J. Acoust. Soc. Am. **95**, 1559–1569.
Cunningham, S. P. (**2003**). "Modelling the recognition of band-pass filtered speech," Doctoral thesis, Department of Computer Science, The University of Sheffield.
Darwin, C. J., and Carlyon, R. P. (**1995**). "Auditory grouping," in *The Handbook of Perception and Cognition, Vol. 6, Hearing*, edited by B. C. J. Moore (Academic, New York), pp. 387–424.
de Cheveigné, A., and Kawahara, H. (**1999**). "Missing-data model of vowel identification," J. Acoust. Soc. Am. **105**, 3497–3508.
Drullman, R. (**1995**). "Speech intelligibility in noise: relative contributions of speech elements above and below the noise level," J. Acoust. Soc. Am. **98**, 1796–1798.
Drygajlo, A., and El-Maliki, M. (**1998**). "Speaker verification in noisy environment with combined spectral subtraction and missing data theory," *Proc. ICASSP-98*, pp. 121–124.
Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.
Fletcher, H. (**1953**). *Speech and Hearing in Communication* (Van Nostrand, New York).
Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.
Gales, M. J. F., and Young, S. J. (**1993**). "HMM recognition in noise using parallel model combination," *Proc. EUROSPEECH*, pp. 837–840.
Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (**1992**). "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," NIST.
Ghitza, O. (**1993**). "Adequacy of auditory models to predict human internal representation of speech sounds," J. Acoust. Soc. Am. **93**, 2160–2171.
Gustafsson, H. A., and Arlinger, S. D. (**1994**). "Masking of speech by amplitude-modulated noise," J. Acoust. Soc. Am. **95**, 518–529.
Holube, I., and Kollmeier, B. (**1996**). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am. **100**, 1703–1716.
Howard-Jones, P. A., and Rosen, S. (**1993**). "Uncomodulated glimpsing in 'checkerboard' noise," J. Acoust. Soc. Am. **93**, 2915–2922.
Hyvärinen, A., Karhunen, J., and Oja, E. (**2001**). *Independent Component Analysis* (Wiley, New York).
Kasturi, K., Loizou, P. C., Dorman, M., and Spahr, T. (**2002**). "The intelligibility of speech with 'holes' in the spectrum," J. Acoust. Soc. Am. **112**, 1102–1111.
Kryter, K. D. (**1962**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–1697.
Lippmann, R. P. (**1996**). "Accurate consonant perception without mid-frequency speech energy," IEEE Trans. Speech Audio Process. **4**, 66–69.
Miller, G. A. (**1947**). "The masking of speech," Psychol. Bull. **44**, 105–129.
Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.
Miller, G. A., and Nicely, P. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Moore, B. C. J. (**2003**). "Temporal integration and context effects in hearing," J. Phonetics **31**, 563–574.
Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**). "A model for the prediction of thresholds, loudness, and partial loudness," J. Audio Eng. Soc. **45**, 224–240.
Moore, B. C. J., Glasberg, B. R., Plack, C. J., and Biswas, A. K. (**1988**). "The shape of the ear's temporal window," J. Acoust. Soc. Am. **83**, 1102–1116.
Palomäki, K. J., Brown, G. J., and Barker, J. (**2002**). "Missing data speech recognition in reverberant conditions," *Proc. ICASSP*, pp. 65–68.
Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (**1988**). "SVOS Final Report: The Auditory Filterbank," Technical Report 2341, MRC Applied Psychology Unit.
Raj, B., Singh, R., and Stern, M. (**1998**). "Inference of missing spectrographic features for robust speech recognition," *Proc. ICSLP*, pp. 1491–1494.
Roman, N., Wang, D., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**, 2236–2252.
Scott, S. K., Rosen, S., Wickham, L., and Wise, R. J. S. (**2004**). "A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception," J. Acoust. Soc. Am. **115**, 813–821.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Shannon, R. V., Jensvold, A., Padilla, M., Robert, M. E., and Wang, X. (**1999**). "Consonant recordings for speech testing," J. Acoust. Soc. Am. **106**, L71–L74.
Simpson, S., and Cooke, M. P. (**2005**). "Consonant identification in *N*-talker babble is a nonmonotonic function of *N*," J. Acoust. Soc. Am. **118**, 2775–2778.
Sinex, D. G., and Geisler, C. D. (**1983**). "Responses of auditory-nerve fibres to consonant-vowel syllables," J. Acoust. Soc. Am. **73**, 602–615.
Steeneken, H. J. M., and Houtgast, T. (**1979**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.
Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). "Dynamic specification of coarticulated vowels," J. Acoust. Soc. Am. **74**, 695–705.
Tchorz, J., and Kollmeier, B. (**2002**). "Estimation of the signal-to-noise ratio with amplitude modulation spectrograms," Speech Commun. **38**, 1–13.
Varga, A. P., and Moore, R. K. (**1990**). "Hidden Markov model decomposition of speech and noise," *Proc. ICASSP*, pp. 845–848.
Wang, M. D., and Bilger, R. C. (**1973**). "Consonant confusions in noise: a study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.
Warren, R. M. (**1970**). "Perceptual restoration of missing speech sounds," Science **167**, 392–393.
Warren, R. M., Riener, K. R., Bashford, J. A., and Brubaker, B. S. (**1995**). "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," Percept. Psychophys. **57**, 175–182.
Warren, R. M., Hainsworth, K. R., Brubaker, B. S., Bashford, J. A., and Healy, E. W. (**1997**). "Spectral restoration of speech: Intelligibility is increased by inserting noise in spectral gaps," Percept. Psychophys. **59**, 275–283.
Young, S. J., and Woodland, P. C. (**1993**). "HTK Version 1.5: User, Reference and Programmer Manual," Cambridge University Engineering Department, Speech Group.

Estimates of effective frequency selectivity based on the detection of a tone added to complex maskers

Virginia M. Richards[a] and Zhongzhou Tang
Department of Psychology, University of Pennsylvania, 3401 Walnut Street, Suite 302C, Philadelphia, Pennsylvania 19104

(Received 24 January 2005; revised 29 November 2005; accepted 13 December 2005)

In Experiment 1, the validity of parameters associated with the roex(p,r) auditory filter shape was examined for three different types of maskers: (a) A noise masker, (b) a random 12-tone masker whose frequencies varied on a burst-by-burst basis [multiple-burst different (MBD)], and (c) a random 12-tone masker whose frequencies were the same across bursts [multiple-burst same (MBS)]. First, the power spectrum model of masking was used to estimate auditory filter shapes for four observers. Second, the resulting auditory filter shapes were used in a computer simulation that provided an estimate of internal noise for each observer. Third, relative weights across frequency were estimated for each observer and each masker type. For the noise masker, these analyses provided predictions and relative weights that were consistent across the three analyses. For the MBD and MBS maskers, there was little consistency; neither the estimated internal noise nor the estimated relative weights reliably supported a single-filter model of detection. In Experiment 2, the time course for the detection of a tone added to an MBD masker was evaluated by estimating relative weights jointly in time and frequency. The relative weights at the signal frequency formed a rough inverse "U" across time. [DOI: 10.1121/1.2165001]

PACS number(s): 43.66.Dc, 43.66.Mk [RAL] Pages: 1574–1584

I. INTRODUCTION

The detectability of a fixed-frequency tone added to a random multitone masker can be very poor, particularly when the number of components comprising the masker is small (Neff and Callaghan, 1988; Oh and Lutfi, 1998). The impact of this "informational masking" is largest when the frequencies of the masker components are drawn at random prior to each stimulus presentation compared to, for example, each two-interval forced-choice trial. Following the work of Leek *et al.* (1991) and Neff *et al.* (1993) provided estimates of the shape of "attentional" filters, that is, the effective and integration region for an assortment of notched maskers, including a noise masker and a random ten-tone masker. Here, the term auditory filter will be used in place of Neff *et al.*'s "attentional filter." Neff *et al.* used the power spectrum model of masking (PSMM) and a roex(p,r) weighting function with efficiency parameter K (Patterson *et al.*, 1982; Patterson and Moore, 1986) to estimate the filter shapes. For the notched-noise maskers, thresholds were relatively similar across observers. In contrast, there were large individual differences when the random multitone maskers were tested (Neff *et al.*, 1993). Even though thresholds differed for these two types of maskers, the estimated bandwidths of the auditory filters were similar. The efficiency parameter, however, varied across observers with the differences being amplified when random multitone maskers were used. The very low efficiencies estimated for some observers in the random multitone masker condition might reflect any of a number of possible factors; general uncertainty effects, the deleterious effects of integration of energy at regions remote from the signal (e.g., Neff and Callaghan, 1988; Oh and Lutfi, 1998; Richards *et al.*, 2002), etc. Whatever the reason for the change in efficiency across masker conditions, it is notable that the application of the PSMM led to similar bandwidth estimates for both the noise and random multitone maskers.

Kidd and colleagues (e.g., Kidd *et al.*, 1994) have demonstrated that the informational masking associated with a tone added to random multitone maskers can be substantially reduced if the masker is random in both time and frequency. Consider two types of maskers, the multiple-bursts-different (MBD) and multiple-bursts-same (MBS) maskers (Kidd *et al.*, 1994). Drawing just one example from Kidd *et al.* (1994), the maskers were composed of eight sequential bursts, each burst being composed of eight tone pips. The signal to be detected was a coherent series of eight, 1000 Hz tone pips. For the MBD masker, the frequencies of the eight tones were randomly chosen for each burst. For the MBS masker, the frequencies were chosen at random for the first burst, but the same frequencies were used for the subsequent seven bursts. This masker is similar to the random multitone masker used by Neff *et al.* (1993), the difference being that the stimuli are comprised of bursts. Kidd *et al.* (1994) found that detection thresholds for the MBD maskers were very low, and on average 40 dB below the MBS thresholds. This release from informational masking was attributed to the fact that when the masker is random in time and frequency, the fixed-frequency signal "stream" stands out from the masker and is therefore easier to detect (see also Kidd *et al.*, 1995).

One aim of the first experiment was to estimate the auditory filter shape for the MBD masker made popular by Kidd and colleagues. Noise and MBS maskers were also

[a]Author to whom correspondence should be addressed; electronic mail: richards@psych.upenn.edu

tested and evaluated, allowing a comparison of filter estimates for all three types of masker. The methods used in Experiment 1 mirror those used by Neff *et al.* (1993). For each masker type, the PSMM was used to estimate the shape of the effective filter weighting function using a roex(p,r) shape, along with the efficiency parameter K (Patterson *et al.*, 1982; Patterson and Moore, 1986).

A second aim of Experiment 1 was to corroborate the derived auditory filter shape using alternative methods. This was done because psychophysical data analyzed using the procedures described by Patterson *et al.* (1982) do not take into account the trial-by-trial variation in the power of the stimuli. In consideration of this issue, Patterson and Henning (1977) showed that for roex weighting functions, there was little cost of ignoring variability for *noise* maskers. It is not obvious whether their analysis would generalize to the highly variable MBD and MBS maskers. To answer this question, computer simulations were run. The stimuli were passed through the fitted filters, and the magnitude of internal noise was altered to best predict obtained thresholds. To the degree the thresholds could be well predicted for all masker types, the result would support the estimated auditory filter shapes. If the thresholds could not be predicted, the result would indicate the estimated filter shape might be inappropriate.

In a third component of Experiment 1, relative weights as a function of frequency were estimated using the stimuli and responses obtained when there was no notch in the masker. The relative weights were derived using a logistic regression. The resulting relative weights provide a second corroboration of the estimated filter shape. For example, if the estimated auditory filter bandwidth is wide, yet the relative weights are near zero for all but the signal frequency, the consistency of the estimate of the auditory filter would be questioned.

Past work suggests that for the detection of a tone added to noise, the range of frequencies for which relative weights are nonzero is roughly consistent with the range expected based on estimates of auditory filters (e.g., Ahumada and Lovell, 1971; Ahumada *et al.*, 1975; Gilkey and Robinson, 1986; Richards and Buss, 1996). For multitone maskers similar to the MBS masker, the estimates of relative weights tend to be positive in the region of the signal frequency (Tang and Richards, 2003; Alexander and Lutfi, 2004), although the pattern of weights obtained away from the signal frequency are not wholly consistent across studies (e.g., Alexander and Lutfi, 2004; Gilkey *et al.*, 2001; Richards and Neff, 2004; Tang and Richards, 2003). These past studies suggested that relative weights might provide an independent means of estimating frequency selectivity for noise, MBD, and MBS maskers.

In Experiment 2, relative weights in both frequency *and* time were estimated for the MBD masker. Due to the large number of parameters to be estimated (both time and frequency relative weights), the gradation in frequency was relatively sparse. Consequently, the results speak to the issue of how relative weights build throughout the duration of the detection trial. This is of interest because these relative weights might reflect dynamic changes in attention when temporally complex stimuli such as the MBD masker are tested.

II. EXPERIMENT 1

A. Methods

1. Observers

Four normal-hearing observers ranging in age from 20–30 years participated. All observers had thresholds in quiet of 15 dB HL or better (EarScan MP pure-tone audiometer) for octave frequencies from 250 to 8000 Hz except Obs 4, whose threshold in quiet was estimated to be 20 dB HL at 250 Hz (left ear). Observers were paid for their participation. The experiments were conducted with observers seated individually in double-walled sound attenuated booths.

2. Stimuli

The signal and maskers were comprised of ten sequential bursts, each 30 ms long including 5 ms cosine-squared onset-offset ramps. There was no delay between bursts, i.e., as the offset ramp reached zero voltage the ensuing onset ramp began. The signal to be detected was a sequence of 1000 Hz tone pips. The phases were adjusted so that the ten bursts reflected a continuous tone that was turned on and off using cosine-squared ramps. In the noise condition, the noise ranged in frequency from 333 to 3000 Hz, with an average spectrum level of 45 dB SPL, leading to a total power of 79.3 dB SPL when there was no notch. In the MBD condition, the frequency composition of the masker changed from burst to burst. When there was no notch, each masker burst was the sum of 12 tone pips with random phases and frequencies drawn from a linear frequency scale ranging from 333 to 3000 Hz. When a notch was present, the range of frequencies from which the tone pip frequencies could be chosen was restricted in that the frequencies within the notch were excluded. In the MBS condition, the frequency composition of the first 12-tone burst was chosen as for the MBD masker, and those frequencies repeated for all bursts. The levels in the MBD and MBS maskers were equal to those of the noise maskers, so as the notch width widened the power per component was reduced to maintain an overall level equal to that of the noise stimulus with a constant spectrum level (a change of at most 1.3 dB across all notch widths tested). The tone pips were generated using a fundamental frequency of 33.3 Hz, and the 12 component tones were forced to be of unique frequencies (except a masker component and the signal component could, by chance, share the signal's frequency when there was no notch).

The notch bandwidth surrounding the signal frequency was systematically varied. For the noise maskers, there was either no notch or deviations from the signal frequency (half-notch bandwidths) of 20, 50, 150, 250, and 350 Hz. For the MBD and MBS maskers, a 33.3 Hz fundamental was used, and so deviations from the signal frequency were 33, 67, 167, 267, and 367 Hz. Even so, values of 20, 50 150, 250, and 350 will be used universally in denoting the notch deviations. The stimuli were digitally generated and presented through two channels of a 16-bit DAC using a sampling rate of 20000 Hz, and low-pass filtered at 7000 Hz using

matched filters (KEMO VBF8). The time delay between the two intervals of the two-interval forced-choice procedure was 350 ms. Stimuli were presented diotically over Sennheiser HD410SL headphones initially calibrated using an artificial ear and dichotic loudness matching with calibrated TDH headphones mounted in circumaural cushions (MR-41).

3. Experimental procedures

A two-interval forced-choice procedure with a 2-down 1-up adaptive algorithm was used to estimate the 71% point on the psychometric function (Levitt, 1971). The adaptive algorithm initially altered the signal level by 4 dB and the step size was reduced to 2 dB after three turnabouts. The adaptive procedure terminated after 50 trials and the threshold was the average of the last even number of reversals except the first three. The initial signal level was approximately 10 dB above the ultimate threshold. The signal was as likely to be in the first as the second interval. Observers indicated the presence of a signal with a keyboard press and received visual correct-answer feedback.

Observers 1 and 3 ran the three conditions in the same order; noise followed by MBD followed by MBS. Observer 2 ran the MBD, then noise, then MBS conditions. Observer 4 ran the MBS, then MBD, then noise conditions. In each condition, the notch bandwidths tested were run in either ascending or descending order, depending on whether the sequences started with no notch or a 350 Hz notch, respectively. In each condition, two observers ran ascending sequences and two ran descending sequences. Observers ran at least ten. 50-trial practice sets prior to data collection for each condition/notch bandwidth tested. Thresholds are based on the average of the last 15 threshold estimates.

Relative weights were estimated using a larger number of trials than the threshold estimates used. The final 20, 50-trial sets were used except that very low and very high signal levels were excluded from the analysis (lower and upper 10% of all signal levels across all trials).

4. Computational procedures

a. Power spectrum model of masking. The psychophysical data were fitted to a two-parameter roex(p,r) weighting function and efficiency parameter K using the procedures outlined by Patterson *et al.* (1982). The parameters p and r were fitted using MATLAB'S (The MathWorks, Inc., 2000) *fminsearch* procedure. Similar to the method used by Neff *et al.* (1993), the MBD and MBS maskers were modeled using a uniform probability density function (i.e., the 33.3 Hz fundamental frequency was not taken into account). Note that as the notch bandwidth increased, the probability density function associated with the MBD and MBS maskers changed; that is, the likelihood that one of the 12 masker pips fell at any one frequency increased as the notch bandwidth increased.

b. Simulations. After estimating the roex(p,r) weighting functions, the filter shapes were evaluated using a computer simulation. The procedure was as follows. In the first step, the squared FFT of 10000 maskers (per notch bandwidth) were multiplied by the roex filter shape, and converted to a dB scale. The maskers were generated using the same algorithm as in the experiment, i.e., the "actual" bandwidths were used. Second, a standard deviation of an internal noise, in dB, was chosen. Third, for each notch, the signal level that led to 71% correct signal detections was estimated using a Nelder–Mead simplex method (*fminsearch* in MATLAB). This yielded a threshold estimate for each notch bandwidth for the chosen internal noise. Fourth, the RMS error between the predicted and actual thresholds was determined. The procedure was repeated for several different internal noise standard deviations. Across the different internal noises, a grid search led to an estimate of the smallest error. The result was an estimate of: (a) the variance associated with the internal noise and (b) threshold predictions for each notch bandwidth. Although the procedure has stochastic components, variability in the solutions was found to be small.

c. Relative weights. In the third analysis, linear relative weights in frequency were estimated. Because weights near the signal frequency were of particular interest, only the no-notch stimuli and responses were used. For the noise stimuli, the relative weights were estimated as follows. The power spectra for the stimuli presented in the first and second intervals were determined using an FFT. The power within 33.3 Hz bins was determined, and those values transformed to a dB scale. The resulting power spectra for the first and second interval were differenced. Then a logistic regression was applied to estimate coefficients for each frequency bin (e.g., Alexander and Lutfi, 2004). The normalized regression coefficients were taken as the relative weights. The regression model was fitted using MATLAB'S *glmfit* procedure.

For the MBS and MBD maskers, the procedure was somewhat different. For each burst the frequencies of the tone pips were generated using a 33.3 Hz fundamental, and so the 33.3 Hz binning was present in the stimuli. For the MBS maskers, a vector of 0's and 1's was generated; a 1 indicated energy present at that frequency and a 0 indicated no energy at that frequency. Note that a value of 1 was used at the signal frequency regardless of the signal magnitude and regardless of whether or not a masker component also happened to fall at the signal frequency. The stimulus vectors were differenced across intervals, and a logistic regression between the differenced vectors and observers' responses was conducted.

For the MBD maskers, the vector entries were the number of tone pips that occurred at that frequency (i.e., count across time). The signal, when present, was arbitrarily assigned a value of 3.[1] As for the other conditions, the vectors were differenced across intervals, and logistic regression coefficients taken as relative weights.

After the coefficients had been estimated for each frequency and observer, the weight at the signal frequency was assigned a value of 1, and the other weights scaled accordingly. Finally, the scaled weights were averaged across observers to provide an "average" weighting pattern and associated standard errors of the mean.

For relative weights, a statistical significance criterion of $p<0.01$ was adopted. While $p<0.01$ might be considered conservative for any one weight estimate, given the large number of weights estimated, it is liberal in terms of the reliability of the overall pattern of weights.[2]

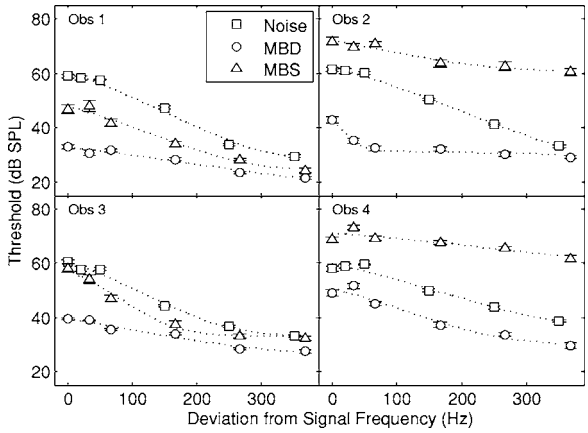


FIG. 1. For each observer thresholds are plotted as a function of the deviation between the edge of the notch and the signal frequency (one-half notch bandwidth). Squares, circles, and triangles are for the noise, MBD and MBS conditions, respectively. Error bars, where visible, indicate the standard error of the mean across fifteen threshold replicates. The dotted curve shows the prediction of the roex(p,r) weighting function fitted to the data sets.

B. Results and discussion

1. Threshold data and fits using the power spectrum model of masking

Figure 1 shows the thresholds in dB SPL as a function of the frequency separation between the signal frequency and the edge of the notch (half-notch bandwidth) for each observer. Results for the noise, MBD and MBS conditions are indicated using squares, circles, and triangles, respectively. The slight difference in actual notch bandwidth for the MBD and MBS compared to the noise maskers is apparent in the plot (squares slightly to the left of the other symbols). Error bars indicate the standard errors of the mean across 15 replicate threshold estimates. The dotted curves shows the predictions from the PSMM and the roex(p,r) weighting function fitted to the data sets (Patterson *et al.*, 1982).

For the noise and MBD maskers, the pattern of thresholds is similar for all four observers. Thresholds fall as the notch bandwidth increases, and the rate of threshold change with notch bandwidth is lower for the MBD than the noise condition. For the MBS condition, thresholds are different for Obs 1 and 3 (thresholds lower for the MBS than the noise masker; triangles versus squares) compared to Obs 2 and 4 (thresholds higher for the MBS than the noise masker; triangles versus squares). Additionally, the rate of change in threshold with notch bandwidth is shallower for Obs 2 and 4 compared to Obs 1 and 3. Presumably Obs 1 and 3 would fall into Neff *et al.*'s (1993) "low threshold" category whereas Obs 2 and 4 would be categorized as "high threshold" listeners.

Table I lists the fitted K (efficiency in dB), r (dynamic range/asymptote in dB), p (bandwidth parameter), equivalent rectangular bandwidths (ERB), and percent of variance accounted for by the fit (PCVar) for each observer and masker type. The threshold data are well fitted by the PSMM. In the noise condition, the Ks are smaller and the rs larger than typical (e.g., Wright, 1996; Neff *et al.*, 1993), potentially due to a variety of differences in the current and past experiments (e.g., different signal frequency compared to Wright (1996), stimuli that are presented using sequences of bursts, *etc.*). Because the MBD and MBS stimuli were modeled using a scalar of the probability density function, an arbitrary choice, there is little to be gained by comparing relative values of the K and r across these conditions, nor against values found elsewhere (e.g., Neff *et al.*, 1993).

TABLE I. The fitted parameters K (efficiency in dB), r (dynamic range), p, and ERB (equivalent rectangular bandwidth) and PCVar (percent of variance accounted for by the model) for each observer in each masker condition.

	Observer	K (dB)	r(dB)	p	ERB(Hz)	PCVar
Noise	1	−5.9	−44	30.3	132	99
	2	−4.8	−45	24.9	161	100
	3	−5.4	−39	32.3	124	99
	4	−8.0	−34	20.7	193	98
MBD	1	−37.8	−23	11.4	351	97
	2	−18.8	−29	92.6	43	98
	3	−29.9	−22	15.8	254	96
	4	−16.4	−32	25.0	160	96
MBS	1	−19.1	−35	25.8	155	98
	2	3.4	−21	18.9	211	95
	3	−7.2	−38	43.1	93	99
	4	−0.9	−100	7.9	507	81

Finally, consider the values of p and the associated ERBs. Recall that if the function relating threshold and notch bandwidth has a consistently flat slope, one would expect a small p, or large ERB. This holds for the current data; as expected based on Fig. 1, the ERBs for Obs 2 and 4 are wider in the MBS condition than the other conditions. By a similar argument, the wider ERBs estimated in the MBD condition for Obs 1 and 3 are expected. Less obvious is the very narrow ERB (43 Hz) estimated for Obs 2 with the MBD masker. For the noise masker, the ERBs are close to what would be expected based on past results (e.g., Neff *et al.*, 1993; Moore, 1987; Wright, 1996).

To summarize, the PSMM and the roex(p,r) weighting functions provide good fits to the data summarized in Fig. 1. However, there are reasons to question the usefulness of the estimated filter shape for the MBD and MBS data sets. First, values of the ERBs are extreme. Both Moore (1987) and Wright (1996) have reported distributions of ERB for 2000 Hz tones added to noise. Adjusting their ERBs downward (the current signal is 1000 Hz), for the MBD and MBS maskers, three quarter of the current ERBs are outside the range reported by Moore (1987) and half are outside the range reported by Wright (1996). Second, some of the thresholds are very high, leading to very large values of K. These extreme parameter estimates suggest that although the data are well fitted using the model, the underlying model may not be appropriate when non-noise maskers are tested. Following procedures similar to those used by Patterson and Patterson and Henning (1977), a computer simulation was run to predict thresholds for the noise, MBD and MBS maskers.

2. Simulations

For these simulations the stimuli were initially passed through the roex(p,r) filter previously estimated. Best fits to

thresholds were estimated by varying the amount of "internal noise" added to the filter output. The results are quite clear: (a) The noise data were well predicted by the simulated model, and (b) the simulations could not account for thresholds in the MBD and MBS conditions.

For the noise maskers, the standard deviation of the internal noise ranged, across listeners, from 1 to 1.8 dB. The predicted thresholds accounted for 98% or more of the variance available to be accounted for (akin to the values of PCVar shown in Table I).

For the MBD condition, the mean threshold always provided a better fit to the data than did the simulation predictions. The actual thresholds were lower than could be predicted even when no internal noise was added to the filter output. This is because the large variance in power at the filter output swamped the change in energy associated with the added signal.

For the MBS condition and for Obs 1, the situation mirrored the situation in the MBD condition—thresholds in the MBS condition were too low to be described using the estimated roex filter weighting functions. For Obs 4, the situation was the opposite—thresholds were too high to be accounted for. An internal noise with a standard deviation over 10 dB was needed to increase predictions to match Obs 4's high thresholds, and even then the mean threshold provided a better fit to the data than did the simulation. For Obs 2 and 3, an internal noise with standard deviations of 7.2 and 1 (the value of 1 being similar to the value obtained in the noise condition) dB were required to model 66% and 45% of the variance available to be accounted for, respectively.

To summarize, for the noise condition the simulations provided good fits to the data. For the MBD and MBS conditions, the simulated model failed to predict thresholds. In the low-threshold MBD condition, in order to maintain a single-filter model, the ERB would have to be very narrow. These results are consistent with the obtained extreme parameter values provided by the PSMM in pointing out the need for care when evaluating MBD and MBS maskers using the traditional PSMM and roex(p,r) weighting function.

3. Relative weights

Figure 2 shows normalized coefficients, or relative weights, averaged across observers for the noise (top), MBD (middle), and MBS (bottom) maskers. Recall the relative weights were derived from the no-notch conditions. The asterisks shown in the bottom of each panel indicate frequencies at which the magnitudes of the normalized coefficients are more than 2.6 standard errors of the mean (across observers after normalizing coefficients, $p<0.01$) distant from zero. The error bars to the left and right of each function are the 2.6 times the standard errors of the mean over observers then averaged across frequency.

For the noise masker (top panel), the individual weighting functions are similar to the averaged function. The averaged function indicates a positive peak at the signal frequency and negative weights just below the signal frequency. The presence of both positive and negative weights in the region of the signal frequency is consistent with patterns derived using a variety of linear weight estimation methods

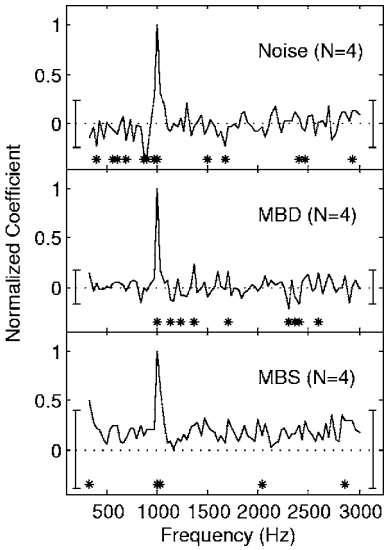


FIG. 2. Averaged normalized coefficients are plotted as a function of frequency. Asterisks indicate the coefficients separated from zero by at least 2.6 ($p<0.01$) standard errors. The error bars to the left and right were computed as follows: First 2.6 times the standard error of the mean across observers was estimated for each frequency bin, and then estimates were averaged across frequency.

(Ahumada and Lovell, 1971; Ahumada *et al.*, 1975; Gilkey and Robinson, 1986; Richards and Buss, 1996). The results also indicate nonzero weights distant from the signal frequency (asterisks). Note the sequence of statistically significant weights in the region of the signal frequency. This sequence (two two-touples) extends across a range of 166 Hz, which is roughly similar to the ERBs indicated in Table I. Thus, even though the significant negative weights in the region of the signal indicate that the PSMM is probably a poor model of detection, the relative weights do not provide compelling evidence that the integration bandwidth in the region of the signal frequency is any different from those estimated using the PSMM.

The middle panel shows the averaged weighting functions for the MBD condition. This function is similar to those estimated for the individual observers. Like the situation with the noise masker, energy distant from the signal frequency is sometimes correlated with observers' detection decisions. More notable is the absence of significant weights in the immediate frequency region of the signal. The absence of statistically significant weights below the signal frequency is especially compelling in this regard. These results suggest, but do not assure, that MBD thresholds reflect narrower filters than the noise thresholds. Moreover, both the pattern of thresholds (Fig. 1) and the pattern of relative weights suggest that the detection mechanisms associated with the detection of a tone added to noise and a tone added to a MBD masker may be quite different.

For the MBS maskers the relative weights varied markedly across observers. Therefore, in addition to the averaged pattern shown in the bottom panel of Fig. 2, Fig. 3 shows the relative weights estimated for each of the observers (indicated to the left; O1–O4). As apparent in the averaged pattern of Fig. 2, the relative weight at the signal frequency tended to be maximally positive. The primary difference

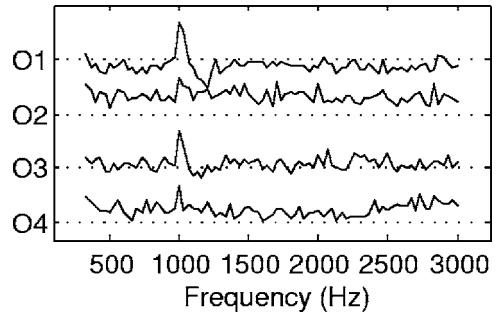


FIG. 3. Normalized coefficients are plotted as a function of frequency for each observer (O1–O4) in a separate panel. The dotted line indicates zero.

across observers is that for two observers (high-threshold Obs. 2 and 4) all relative weights are larger than zero. For the other two observers, relative weights tend to be near zero (Obs. 3) or negative (Obs. 1). Individual differences in the pattern of relative weights led to averaged relative weights greater than zero for all frequencies (Fig. 2). The individual differences in the patterns of relative weights are not surprising given the large individual differences in detection thresholds for the MBS masker. Overall, the variation in individual relative weights prevents summary statements for the MBS masker.

4. Summary of Experiment 1

Estimates of frequency selectivity derived using the PSMM and the notched-masker method are not assured to be meaningful when evaluating the detection of a tone added to complex maskers such as MBD and MBS maskers. This is apparent in the fact that good fits to a notched-masker detection experiment could be achieved using the PSMM even though simulations using the estimated filters could not predict thresholds. Presumably this is because the PSMM does not take into account the large intertrial variability in power associated with the MBD and MBS maskers. Nonetheless, this result does not suggest that the simulated energy model is any better a model of detection than the original PSMM model; the existence of negative relative weights argues against power/energy models of detection in general.

An analysis of relative weights suggests that for the averaged data single-filter estimates of frequency selectivity for the MBD condition would be quite narrow. A tenuous working hypothesis is that this result is tied to the percept associated with the MBD masker, namely, that the signal appears to form a stream that is perceived as separate from the masker (Kidd *et al.*, 1995).

The presence of strictly positive weights for the high-threshold listeners in the MBS condition may be interpreted in terms of one of the conclusions offered by Neff *et al.* (1993); that when a tone is masked by random multitone maskers, observers with high and low thresholds differ in their relative efficiencies, not necessarily the bandwidth of the underlying auditory filters. By this interpretation, inefficiency would reflect the inappropriate integration of energy from frequency regions distant from the signal (e.g., Neff and Callaghan, 1988; Oh and Lutfi, 1998; Richards *et al.*, 2002), i.e., positive weights at nonsignal frequencies.

III. EXPERIMENT 2: TIME-BY-FREQUENCY RELATIVE WEIGHTS FOR AN MBD-LIKE STIMULUS

In this experiment, relative weights were determined in both time and frequency for a complex stimulus similar to the MBD masker used in Experiment 1. The task was to detect a coherent sequence of 1000 Hz tone pips amongst other randomly varying tone-pip maskers. In this exploratory experiment, the aim was to examine the time course of the formation of the signal stream separate from the masker.

The starting point of stimulus construction was a time-by-frequency matrix. The no-signal stimulus was constructed by placing a tone pip in each time-by-frequency cell with probability p_t (the probability a tone fell in any one cell of the no-signal stimulus). For the signal stimulus, a sequence of 1000 Hz tone pips was enforced. Additionally, for the signal stimuli the probability that a tone pip fell at a nonsignal frequency was reduced so that the total expected number of tone pips in any stimulus, signal or no-signal, was the same. This experimental design is similar to that tested by Kidd *et al.* (1995) and the MBD condition of Experiment 1. The dominant difference between the current experiment and those of Kidd *et al.* (1995) is that in the current experiment the signal frequency was fixed rather than randomly drawn.

The independent variable for this task was p_t, the probability that a tone pip falls in any one cell for the no-signal stimulus. For large p_ts, the task was expected to be difficult; if many of the cells have tone pips, sequences of tone pips would be likely to occur at the signal frequency whether the signal was present or not. For small p_ts, the task would be easier because sequences of tone pips at the signal frequency would rarely occur in no-signal stimuli. Our expectation was that observers' ability to detect the coherent signal would depend monotonically on p_t. To evaluate this assumption, psychometric functions relating p_t and d' were obtained.

While there are many masking studies that have provided relative weights as a function of frequency, only a few studies have estimated joint time-by-frequency weighting functions. Ahumada *et al.* (1975) estimated relative weights in time and frequency for the detection of a tone added to noise. The tone was shorter than the noise and situated in the temporal center of the noise. Considering only responses to the signal trials, Ahumada *et al.* found negative weights at the signal frequency just prior to the signal onset and just after the signal offset. Gilkey and Robinson (1986), whose signal was shorter than the noise masker, additionally found that for the signal frequency relative weights were initially small or negative, grew over time, and then declined. Moreover, Ahumada *et al.* reported that during the course of the signal, negative weights were obtained at frequencies just above and just below the signal frequency. For the no-signal trials, negative weights were not obtained (see also Ahumada and Lovell, 1971). Ahumada *et al.* suggested the negative weights obtained for the signal trials reflected a contrast—observers compared the signal against the adjacent frequencies and adjacent temporal epochs. Weighting functions for the noise masker in Experiment 1, which were estimated using both signal and no-signal intervals, indicated negative coefficients for frequencies just below, but not just above, the signal frequency.

An additional study in which both time and frequency weights were obtained is a profile analysis task evaluated by Dai and Berg (1992). The observer's task was to detect an increment in level at the signal frequency compared to a background of equal-level tones. Amplitude modulated tones were tested in which there were three epochs. Thus, an increment might occur at the beginning, middle, or end, epoch at the signal frequency. Their results indicate that, for sufficiently long stimuli, observers exhibited positive weights at the signal frequency and negative weights at the non-signal frequencies. Moreover, and again for sufficiently long stimuli, observers adjusted their weights, and perhaps their attention, when just one temporal epoch of the signal was elevated in level. For this profile analysis study, which included overall level randomization, the contrast between the signal and nonsignal components dominated the obtained weighting patterns. For the MBD-like task examined here, it is not obvious, *a priori*, whether the perceptual streaming of the signal would likewise be indicated as a contrast (i.e., negative weights at nonsignal frequencies). Even if that description is apt, given the large number of off-signal-frequency relative weights the magnitude of any one weight would be very small (due to an averaging of a large number of off-frequency cells).

A. Methods

1. Observers

Ten normal-hearing observers, ranging in age from 19–31 years, participated. Observers 1–4 also participated in Experiment 1. Of the new observers, all had thresholds in quiet of 15 dB HL or better for octave frequencies from 250 to 8000 Hz. Observers were paid for their participation. The experiments were conducted with observers seated in double-walled sound attenuated booths.

2. Stimuli

As described above, the starting point of the stimulus generation was the formation of a time-by-frequency grid. For one group of observers, group A, Obs 1–7, the grid contained ten 30 ms time epochs (i.e., ten bursts) and 19 frequency bins. The potential frequencies were separated by one-sixth of an octave and ranged from 333 to 3000 Hz. For the second group of observers, group B, Obs 8–10, the grid included ten 30 ms time epochs and 30 frequencies. The potential frequencies were separated by 1/6.5 octaves and ranged from 200 to 5000 Hz. It seemed unlikely that the difference in parameters across groups would be important; the results for the two groups are included so as to provide a replication of results for this new type of stimulus set.

The dependent variable, p_t, is the probability that a tone pip would be present in any one cell of the no-signal stimulus. For example, for group A, there were 10×19 or 190 possible tone pips and the no-signal stimuli were generated by independently choosing whether or not a tone pip occurred in each cell. Filling the ten temporal epochs at the signal frequency, and then independently choosing whether or not a tone pip occurred at the remaining time-by-frequency locations generated the signal stimulus. For the signal stimuli, the likelihood that a tone occurred at a non-signal frequency was reduced so that the expected number of tone pips was the same for the no-signal and the signal stimuli. Note that if very small values of p_t are tested, this constraint could not be maintained, i.e., for very small p_ts a total of ten tone pips would rarely occur. Each 60 dB SPL tone pip was 30 ms in duration, including 5 ms cosine-squared onset and offset ramps. There was no delay between bursts of tone pips; i.e., as the offset of one tone pip approached zero voltage the onset of the other began. The signal frequencies were 1000 and 950 Hz for the groups A and B, respectively. For group A, the phases were randomly chosen for each presentation, but the phases were consistent across time epochs (e.g., if a series of tone pips had a single frequency, it was as though a single tone was continuous with offset and onset ramps). For group B, the phases were drawn at random for each tone pip, including the signal tone pips. Recall that the signal is defined by virtue of a sequence of equal-frequency tone pips. Thus, the level of the tone pips at the signal frequency was equal to the level of all other tone pips, 60 dB SPL.

The stimuli were digitally generated and presented through a single channel of a 16-bit DAC using a sampling rate of 20000 Hz, and low-pass filtered at 7000 Hz (KEMO VBF8). The stimuli were presented diotically over Sennheiser HD410SL headphones.

3. Experimental procedures

For both portions of this experiment, i.e., estimates of psychometric functions and relative weighting, data was collected in blocks with fixed values of p_t. A yes/no procedure was used.

a. Psychometric functions. Functions relating p_t and d' were measured for Obs 5–7. Data collection was blocked by p_t, starting with small p_ts and increasing to larger p_ts (larger to smaller d's). In most instances, at least ten sets of 50 trials were tested, yielding ten estimates of d'. For d's above 3 or so, data was collected using only five, 50 trial sets. After these observers finished data collection for psychometric functions, additional data was collected in order to obtain relative weights.

b. Relative weights. A logistic regression was applied to stimuli and responses in order to estimate the linear coefficients, which were taken as indicating relative weights. For Obs 1, 2, 4, 5, 8–10 all of the data were collected using a single value of p_t. For Obs 3, data obtained from two values of p_t were included in the analysis, and for Obs 6 and 7 the regression was applied to data obtained using three values of p_t. This variation in the protocol allowed the data of Obs 3, 6, and 7 to have large enough numbers of trials for analysis. The goal was to choose a value of p_t that led to a d' of approximately 1.5. This value of p_t was determined during a 2 h practice session. Table II shows the p_ts tested for each observer, except that the mean values are displayed for Obs 3, 6, and 7.

For group A, the logistic regression was fitted using at least 900 stimulus/response pairs for each observer (although for Obs 1 and 5, the number was substantially higher). For group B, the regression was based on more than 1700 stimulus/response pairs per observer.

TABLE II. The values of p_t tested are listed.

Observer	1	2	3	4	5	6	7	8	9	10
p_t	0.6	0.4	0.55	0.3	0.7	0.6	0.5	0.7	0.6	0.5

The logistic regression related observers' binary responses (signal present or absent) to each stimulus token as a function of whether the stimulus token had (1) or did not have (0) energy at each time-by-frequency cell. Error boundaries for the complete time-by-frequency coefficients are 2.6 times the standard errors of the coefficient estimates, i.e., $p < 0.01$.

For groups A and B, 190 and 300 coefficients, respectively, were estimated. Coefficients were separately estimated for the no-signal and signal stimuli as well as for the whole data set for each observer. We evaluated a variety of normalization schemes that would allow an averaging of relative weights across observers. Ultimately, we compared these options against the pattern of coefficients obtained when the stimuli and responses accumulated from all observers and all trials (both signal and no-signal). The latter, global, approach provides coefficients that are more or less a weighted average across observers, where the weights depend on the proportion of trials each observer contributed to the total data set. The variety of approaches provided similar results. The global approach was preferred because there was no need to choose a normalization criterion.

In addition to presenting time-by-frequency coefficients, summary data are further reduced by presenting relative weights for just frequency and just time (at the signal frequency). For the frequency patterns, relative weights were obtained by averaging the global coefficients across time. For this averaged pattern, error boundaries are estimated as 2.6 times the standard error of the mean. The temporal patterns reported below are the coefficients estimated at the signal frequency as a function of time. For the temporal patterns, the error boundaries are 2.6 times the standard errors of the coefficient estimates.

B. Results and discussion

1. Psychometric functions

Figure 4 plots d' as a function of p_t for Obs 5–7. Note that the values of p_t are descending along the abscissa. The relation between d' and p_t is nearly linear, the dynamic range is manageable (i.e., very low values of p_t are not required for large d's), and sensitivity does not vary hugely across observers.

2. Estimated coefficients

Figure 5 shows the coefficients separately for each member of group A using checkerboard plots. The dashed lines indicate breaks between the panels. The scale is shown in the upper right portion of the figure; lighter grid elements indicate more positive coefficients. The observer associated with each panel is indicated using white lettering (O1, O2, etc.). For each observer, both positive and negative coefficients were estimated. Recall that the coefficients were derived using all trials each observer encountered, i.e., both signal and no-signal trials. The results are clear in that there is a tendency for the maximal positive weights (white) to occur at the signal frequency. Additionally for the two observers who ran the most trials, Obs 1 and 5, there is less variation in the off-frequency coefficients, presumably because the larger number of trials provided more stable estimates of the coefficients.

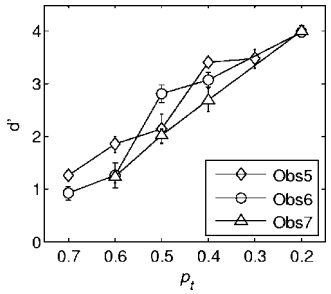


FIG. 4. Psychometric functions (d' as a function of p_t) are shown for Obs 5–7. Error bars are the standard errors of the mean.

Figure 6 shows the summary in which relative weights were estimated using all stimuli and responses for group A. The top panel shows the average weights using a checkerboard plot as in Fig. 5. The bottom left panel shows the coefficients averaged across time and the bottom right panel shows the coefficients at the signal frequency as a function of time. Error bars indicate 2.6 times the standard errors ($p < 0.01$) of the mean (left, across time) and 2.6 times the standard errors about the signal coefficients ($p < 0.01$; right).

The large positive weights at the signal frequency dominate the relative weights shown in the top panel. At the signal frequency, all but the first weight is significantly larger than zero (lower right hand panel). Returning to the checkerboard plot in the top panel, of the 180 nonsignal weights, 24 (13%) are 2.6 standard errors removed from zero, with one-quarter being positive and three-quarters being negative. This far exceeds the 2 (1%) or so that would be expected had the true weights been zero.

The bottom left panel shows the coefficients after aver-

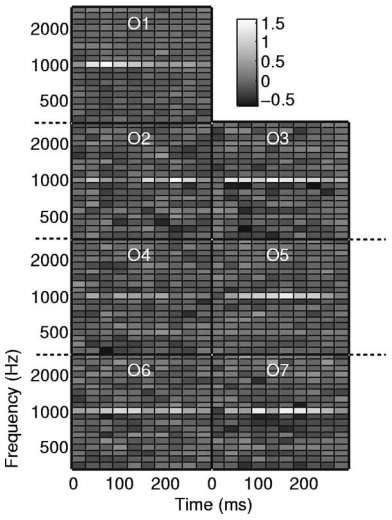


FIG. 5. Checkerboard plots are shown for Obs 1–7. For each plot, the ordinate is frequency (logarithmic scale) and the abscissa is time (linear scale). The brightness of each square indicates the estimated coefficient, as indicated by the scale in the upper right corner. Dashed lines indicate the breaks between panels.

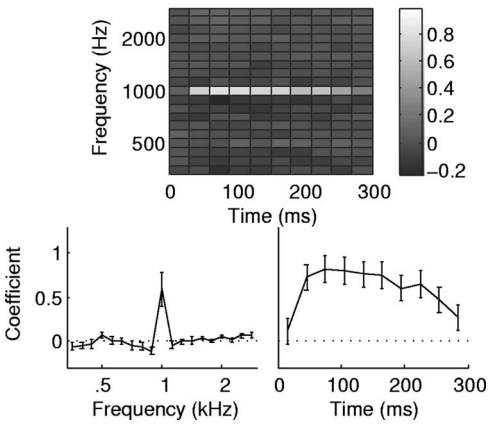


FIG. 6. The top panel shows the checkerboard plot based on all data from all observers of group A. The lower left panel shows the coefficients averaged across time. The lower right panel shows the magnitude of the signal-frequency coefficients as a function of time. Error bars indicate 2.6 times the standard errors of the mean and 2.6 times the standard errors of the estimated coefficients for the lower left and right panels, respectively.

aging across time. The pattern is not too different from that obtained in experiment 1 (Fig. 2). As in the MBD condition of Experiment 1, the peak at the signal frequency dominates the function. Of the 18 nonsignal frequency bins, 11 have coefficients that are removed by at least 2.6 standard errors of the mean from zero. The nonsignal bins were also averaged across frequency and evaluated as a function of time. The resulting function is not plotted because none of ten averaged coefficients were statistically significant.

Figure 7 plots the overall relative weights for group B, the three observers tested using a 30-frequency by 10-time-slot grid. The figure is laid out as Fig. 6. The results are similar to the results of group A. The signal frequency dominates the relative weights (top panel, bottom left panel). The pattern of signal-frequency coefficients as a function of time (bottom right panel) indicates that the first coefficient is not statistically different from zero, consistent with the results for group A. For group B, the temporal function peaks somewhat later in time that for group A; but otherwise the two functions are remarkably similar. Like for group A, when the nonsignal bins were averaged across frequency and evaluated as a function of time, none of 10 averaged coefficients were statistically significant.

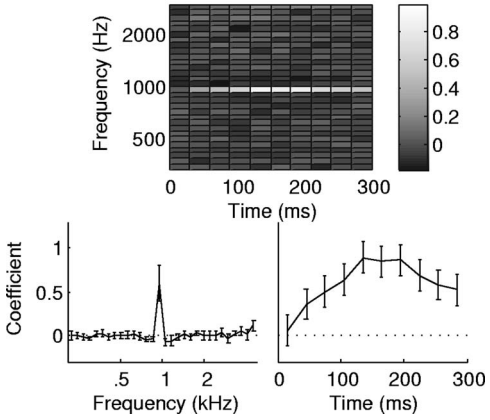


FIG. 7. As Fig. 6, except for group B.

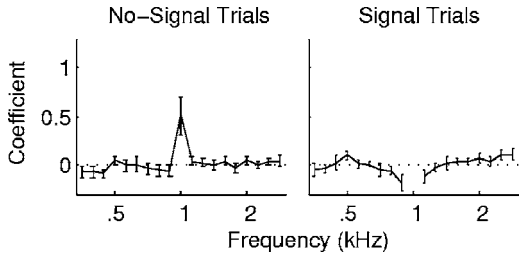


FIG. 8. Coefficients averaged across time are shown separately for regressions applied to just the no-signal or just the signal trials. The results are for group A. In other respects, the plots are like Figs. 6 and 7.

For the frequency weighting patterns (bottom left hand panel), 6 of the 29 nonsignal weights are statistically different from zero, a far smaller proportion of the coefficients than obtained from the data of group A (61% compared to 20%). This difference in the proportion of significant coefficients across groups A and B may reflect a variety of factors; differences in the number of coefficients estimated (190 versus 300 for groups A and B, respectively), differences in the number of trials used to estimate the coefficients (17000 versus 10850 for groups A and B, respectively), etc.

Figure 8 plots the coefficients obtained for group A when the logistic regression was separately applied to the no-signal (left) and signal (right) trials using the data of all observers. For the signal trials, the estimated coefficients are more negative just above and just below the signal frequency than the coefficients estimated using the no-signal trials. (Note that for the signal trials there is no variation in energy at the signal frequency, and so regression coefficients cannot be estimated.) Additionally, for the signal trials but not the no-signal trials, the coefficients grow as the frequency increases for frequencies higher than the signal frequency. The former, but not the latter difference was also apparent for group B. Thus, as described by Ahumada *et al.* (1975, also Ahumada and Lovell, 1971) who studied noise maskers, when the analysis is restricted to just the signal trials, the relative contribution of the frequencies just above and just below the signal frequency *might* be described in terms of a contrast.

Finally, consider the difference in the number of significant nonsignal-frequency coefficients obtained for group A compared to group B, as described when Figs. 6 and 7 were considered. It is noteworthy that the "deviance" (a statistical estimate of quality of fit) is modest for group A and large for group B, suggesting but not assuring that group A's data were better fit than group B's (McCullagh and Nelder, 1999). To additionally evaluate the difference in the proportion of significant nonsignal-frequency weights, we asked whether the full logistic model (all time and frequency cells) provided a significantly better fit to the data than a reduced model in which only the signal-frequency cells contributed to the model. A chi-squared test provides a useful test because the reduced model is nested within the larger model (McCullagh and Nelder, 1999). For both groups of observers the larger model provided a statistically significantly better fit to the psychophysical data than did the reduced model $[\chi^2(180)=579$ and $\chi^2(290)=443; p<0.001]$. This result suggests that energy at frequencies other than the signal frequency contrib-

ute to observers' decisions. Notably, what this analysis does not address is whether the full model better generalizes to a unique data set than would the reduced (signal frequency only) model.

3. Summary of Experiment 2

There are six aspects of the results of Experiment 2 that deserve note. First, coefficients at the signal frequency are elevated relative to the other coefficients. Presumably, this reflects the observers' dependence on the presence of the sequence of bursts at the signal frequency in guiding their decisions.

Second, at the signal frequency, the weight associated with the first burst is not significantly different from zero (temporal relative weights, Figs. 5 and 6). This result was replicated across groups A and B. An examination of the no-signal trials indicates that as the number of signal-frequency sequential tones increased, the false alarm rates also increased. In light of this fact, it is of interest to note that there are more ways a tone pip in, say, the fifth temporal epoch could be one of many sequential tone pips than a tone pip at the first or final temporal epoch. Such probabilistic arguments coupled with the observations that "number of signal-frequency components in a row" correlated with false alarm rates, suggests that smaller coefficients would be obtained for early and late elements of the signal frequency. This prediction of an inverted "U" was not fully realized, meaning that probabilistic arguments do not fully account for the pattern of coefficients at the signal frequency as a function of time. [Potentially similar probabilistic explanations would apply to tasks in which continuous tones are tested, provided it is assumed that the perception of sounds depends on piecewise integration of short segments, or "multiple looks" (Viemeister and Wakefield, 1991)].

Third, for group A, when the coefficients were averaged across time significantly positive weights were obtained at 500 and 2000 Hz, one octave below and above the 1000 Hz signal frequency. This fact is more impressive for the 500 Hz than the 2000 Hz averaged coefficients because for frequencies lower than the signal frequency the weight at 500 Hz was the only one that was significantly different from zero and positive. We tentatively suggest that sequences of 500 Hz and 2000 Hz tone pips may lead to "signal present" responses due to octave confusions associated with pitch estimates.

Fourth, statistical analyses of the data suggest that for both groups A and B, a model based on the linear combination of all time-by-frequency elements is superior to a reduced model in which only the elements at the signal frequency are included. This suggests that, although the impact may be modest, energy at frequencies other than the signal frequency contributed to observers' decisions.

Fifth, consider the difference in the pattern of coefficients obtained using the signal versus the no-signal trials (Fig. 8). For the former set of trials, negative coefficients are apparent at the frequencies just above and just below the signal frequency—for the latter set of trials that result is not obtained. Ahumada *et al.* (1975, also Ahumada and Lovell, 1971) obtained roughly similar results, and suggested the pattern reflects the fact that once observers detected high levels of energy at the signal frequency observers formed contrasts between the signal frequency and neighboring frequencies. An alternative explanation would be that nonlinearities contribute to differences in estimates of linear coefficients (Dai *et al.*, 1996; Richards, 2002). Although the current design was devised, to a large degree, to provide a signal with energy levels equal to those present at the nonsignal frequencies, which in turn ought to enhance the likelihood that the decision processes would be reasonably modeled using a linear model, there are no assurances that the design was successful in this regard. While the implications of this consistency are unclear, it is striking that the data for both noise and MBD maskers share this common feature.

Finally, the pattern of relative weights for the MBD conditions, Experiments 1 and 2 (comparing the middle panel of Fig. 2 and the lower left panel of Figs. 6 and 7) are similar. In both instances, the dominant weight is at the signal frequency. This result suggests there may be little understanding to be gained in estimating all off-frequency weights at each temporal epoch. This is of interest because there are statistical advantages to estimating fewer relative weights.

IV. CONCLUSIONS

The results of the first experiment indicated that single-filter models of detection, such as the PSMM, might not provide reasonable filter weighting function estimates for maskers with substantial variation in time and frequency. This was demonstrated in two steps. First, a filter shape was estimated using typical PSMM methods. Then, the resulting filter was used to process the stimuli. Simulations indicated that thresholds were poorly predicted for the MBD and MBS maskers. In an additional series of measurements, relative weights for the different maskers indicated different patterns of weights for the different maskers. The patterns of weights varied across observers, and so revealed little regarding how listeners integrate information across frequency for the different masker types.

In the second experiment the task was to detect a coherent sequence of 1000 Hz tones against a background of random tone pips. The data from two groups, observers were separately fitted using a logistic linear model. Across groups, there were differences in the proportion of coefficients significantly different from zero, but the pattern of relative weights replicated fairly well across groups. Coefficients indicate large relative weights at the signal frequency and nonzero weights at several other frequencies. Across groups, the off-signal-frequency coefficients, when statistically significant, were as likely to be positive as negative. Collapsed across time, the patterns of relative weights as a function of frequency were similar to those of Experiment 1, where temporal variation was not taken into account. Evaluated as a function of time, the pattern of coefficients at the signal frequency tended to initially increase and then fall.

ACKNOWLEDGMENTS

This work was supported by grant RO1 DC02012 from the National Institutes of Health. Dr. Brian C. J. Moore sug-

gested the simulation approach adopted in Experiment 1. We gratefully acknowledge his contribution. We also thank Christine R. Mason, Associate Editor Dr. Robert A. Lutfi, and two anonymous reviewers for their insightful suggestions on an earlier draft of this manuscript.

[1]The choice of the "signal" size is important because it dictates the magnitude of the coefficient derived at the signal frequency relative to the other frequencies (but not the relative reliability of those weights). This occurs because the "scale" on which the signal was defined (a value of 3) was not the same scale used for the nonsignal time-by-frequency bins (the total number of tone pips). As a result, the magnitude of the coefficient at the signal frequency cannot be directly compared to the coefficients obtained at the nonsignal frequencies. Nonetheless, the *sign* of the signal coefficient, and its statistical significance, is accurate.

[2]A simple computation for the likelihood that at least one of n estimated weights is significantly different from zero is given by $1-(1-p)^n$ where p is the significance level for each estimated weight. This is appropriate whether or not the null hypothesis is true (all weights are zero) and the weight estimates are independent of one another. For Experiment 1, $p=0.01$, and $n=80$, yielding a "pattern-wise" expectation of $1-(1-0.01)^{80}=0.56$. For Experiment 2, the values are 0.85 and 0.95 for groups A and B, respectively. As a result, the precise pattern of weights, per se, for Experiment 2 would not be expected to be fully reproducible. This was deemed acceptable because Experiment 2 is largely exploratory.

Ahumada, A. Jr., and Lovell, J. (**1971**). "Stimulus features in signal detection," J. Acoust. Soc. Am. **49**, 1751–1756.

Ahumada, A. Jr., Marken, R., and Sandusky, A. (**1975**). "Time and frequency analyses of auditory signal detection," J. Acoust. Soc. Am. **57**, 385–390.

Alexander, J. M., and Lutfi, R. A. (**2004**). "Informational masking in hearing-impaired and normal-hearing listeners: sensation level and decision weights," J. Acoust. Soc. Am. **116**, 2234–2247.

Dai, H., and Berg, B. G. (**1992**). "Spectral and temporal weights in spectral-shape discrimination," J. Acoust. Soc. Am. **92**, 1346–1355.

Dai, H., Nguyen, Q., and Green, D. M. (**1996**). "Decision rules of listeners in spectral shape discrimination with or without signal-frequency uncertainty," J. Acoust. Soc. Am. **99**, 2298–2306.

Gilkey, R. H., and Robinson, D. E. (**1986**). "Models of auditory masking: A molecular psychophysical approach," J. Acoust. Soc. Am. **79**, 1499–1510.

Gilkey, R. H., Mason, C. R., and Kidd, G. (**2001**). "Using reproducible noise to illuminate the nature of informational masking," J. Acoust. Soc. Am. **109**, 2468(A).

Kidd, G., Mason, C. R., Deliwala, P. S., Woods, W. S., and Colburn, H. S. (**1994**). "Reducing informational masking by sound segregation," J. Acoust. Soc. Am. **95**, 3475–3480.

Kidd, G. Jr., Mason, C. R., and Dai, H. (**1995**). "Discriminating coherence in spectro-temporal patterns," J. Acoust. Soc. Am. **97**, 3782–3790.

Leek, M. R., Brown, M. E., and Dorman, M. F. (**1991**). "Informational masking and auditory attention," Percept. Psychophys. **50**, 205–214.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 167–177.

McCullagh, P., and Nelder, J. A. (**1999**). *Generalized Linear Models*, 2nd edition (Chapman and Hall/CRC, Boca Raton, FL).

The Math Works, Inc. (**2000**). *Matlab: The Language of Technical Computing* (The Math Works, Inc., Natick, MA).

Moore, B. C. J. (**1987**). "Distribution of auditory-filter bandwidths at 2 kHz in young normal listeners," J. Acoust. Soc. Am. **81**, 1633–1635.

Neff, D. L., and Callaghan, B. P. (**1988**). "Effective properties of multicomponent simultaneous maskers under conditions of uncertainty," J. Acoust. Soc. Am. **83**, 1833–1838.

Neff, D. L., Dethlefs, D. L., and Jesteadt, W. (**1993**). "Informational masking for multicomponent maskers with spectral gaps," J. Acoust. Soc. Am. **94**, 3112–3126.

Oh, E. L., and Lutfi, R. A. (**1998**). "Nonmonotonicity of informational masking," J. Acoust. Soc. Am. **104**, 3489–3499.

Patterson, R. D., and Henning, G. B. (**1977**). "Stimulus variability and auditory filter shape," J. Acoust. Soc. Am. **62**, 649–664.

Patterson, R. D., and Moore, B. C. J. (**1986**). "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing*, edited by B. C. J. Moore (Academic, New York), 123–177.

Patterson, R. D., Nimmo-Smith, I., Weber, D. L., and Milroy, R. (**1982**). "The deterioration of hearing with age: Frequency selectivity, the critical band, the audiogram, and speech thresholds," J. Acoust. Soc. Am. **72**, 1788–1803.

Richards, V. M. (**2002**). "Effects of a limited class of nonlinearities on estimates of relative weights," J. Acoust. Soc. Am. **111**, 1012–1017.

Richards, V. M., and Buss, E. (**1996**). "Frequency correlation functions for the detection of a tone added to modulated noise," J. Acoust. Soc. Am. **99**, 1645–1652.

Richards, V. M., and Neff, D. L. (**2004**). "Cuing effects for informational masking," J. Acoust. Soc. Am. **115**, 289–300.

Richards, V. M., Tang, Z., and Kidd, G. D. Jr. (**2002**). "Informational masking with small set sizes," J. Acoust. Soc. Am. **111**, 1399–1366.

Tang, Z., and Richards, V. M. (**2003**). "Examination of a linear model in an informational masking study," J. Acoust. Soc. Am. **114**, 361–367.

Viemeister, N. F., and Wakefield, G. H. (**1991**). "Temporal integration and multiple looks," J. Acoust. Soc. Am. **90**, 858–865.

Wright, B. A. (**1996**). "Auditory filter asymmetry at 2000 Hz in 80 normal-hearing ears," J. Acoust. Soc. Am. **100**, 1717–1721.

Using monkeys to explore perceptual "loss" versus "learning" models in English and Spanish voice-onset-time perception[a)]

Joan M. Sinnott, Laura A. Powell, and Jazmin Camchong
Comparative Hearing Laboratory, Psychology Department, University of South Alabama, Mobile, Alabama 36688

(Received 27 July 2004; revised 7 December 2005; accepted 8 December 2005)

Cross-language research using synthetic voice-onset-time series over the past 40 years suggests that the Spanish *lead* contrast is acoustically less salient than the English *lag* contrast. This study examined monkey identification of a labial consonant-vowel voice-onset-time (CV VOT) series (−60 to +70 ms) in order to obtain a linguistically unbiased estimate of *lead* versus *lag* salience. Comparisons were made with both English and Spanish adult human listeners. In a classic two-choice identification test, monkey and Spanish functions were quite variable and showed evidence of sensitivity to three types of voicing cues (*lead* versus *simultaneous* versus *lag*). In contrast, English functions were highly categorical and showed sensitivity to only two types of cues (combined *lead/simultaneous* versus *lag*). Next, listeners were explicitly trained via feedback to differentiate stimuli crossing *lead* and *lag* boundaries. Here, monkey and Spanish performance was initially more symmetrical than English performance, with the latter showing reduced sensitivity to the *lead* boundary, but group differences disappeared after extended training. These results provide evidence for *perceptual loss* in English listeners for aspects of Spanish voicing *lead* perception. [DOI: 10.1121/1.2162440]

PACS number(s): 43.66.Gf, 43.71.Es, 43.71.Ft, 43.71.Gv, 43.71.An [ALF] Pages: 1585–1596

I. INTRODUCTION

Pisoni *et al.* (1994) theorized in detail about how such processes as *perceptual loss* and *perceptual learning* could influence the development of human speech perception. According to these authors, "*Universal theory* assumes that newborn infants are capable of discriminating all the possible phonetic contrasts that may be used in any natural language…that infants come equipped with broadly tuned perceptual and attentional mechanisms.…furthermore, the absence of exposure to contrasts that are not phonetically distinctive results in a selective loss of the ability to discriminate those contrasts" (p. 125). In contrast, "*Perceptual learning theory* assumes that the ability to discriminate any particular phonetic contrast is highly dependent on specific early experience with that sound contrast in the language-learning environment" (Pisoni *et al.*, 1994, p. 126). While these authors present an interesting and detailed discussion of the hypothesized mechanisms of *universal-loss* versus *learning*, as well as other mechanisms such as *attunement*, they do not propose any operational definitions of these mechanisms such that they could be accurately measured in an actual experiment.

The goal of this paper is to explore how monkeys can be used (operationally) to investigate perceptual *loss* versus *learning* mechanisms in human cross-language speech perception, specifically with regard to perception of the Spanish voice-onset-time contrast.

[a)]Portions of these data were presented at the Spring 2000 Meeting of the Acoustical Society of America in Atlanta, GA.

A. Background and significance

Voice-onset-time (VOT) has probably received more attention than any other speech feature in cross-language, developmental, and comparative research. Study began with the classic production study of Lisker and Abramson (1964), which demonstrated three basic modes of voicing in consonant-vowel (CV) initial stops: (1) voicing onset precedes the release burst (VOT *lead*); (2) voicing starts more or less simultaneously with the burst; and (3) voicing onset follows the burst (VOT *lag*). These authors further showed that various languages choose different modes of VOT in establishing phoneme categories. For example, English and Spanish are both two-category languages, but English contrasts (2) with (3), while Spanish contrasts (1) with (2). Thai is a three-category language that contrasts all three modes. No languages were reported to contrast (1) with (3). The authors acknowledged that other cues besides VOT (e.g., aspiration noise) contributed to voicing contrasts, but they maintained that these more or less correlated with VOT, so they considered VOT to be the major parameter to investigate perceptually.

1. VOT perception in adults

In their complimentary perception study, Lisker and Abramson (1970) synthesized labial, alveolar, and velar VOT series ranging from −150 to +150 ms. VOT *lead* was produced by adding a low frequency voice bar before a 0-ms VOT CV, and VOT *lag* was produced by cutting back F1 and adding noise during the cutback. They presented the stimuli to English, Spanish, and Thai listeners for identification in terms of their native phonemes. Perception, like production,

appeared language specific. For example, for labials, the English and Thai phoneme boundaries matched up well with the production data. However, Spanish perception did not show a good match, since the production data indicated a boundary in the *lead* range (near −10 ms), but the perception data indicated a boundary shifted to the *lag* region (near +15 ms). The mismatch in the Spanish data was an indication that the *lead* VOT boundary in the synthetic series was acoustically less salient than the *lag* boundary.

Williams (1977a) also studied the perception of English versus Spanish VOT, using both monolingual and bilingual adult listeners. (Bilinguals were defined as having exposure to both languages before the age of six and having no accent in either language.) For monolinguals, the identification tests showed a good match between production and perception, with an English boundary near +25 ms, and a Spanish boundary near −10 ms. But the discrimination tests showed an interesting asymmetry: While the English listeners showed a major peak at the English boundary and no sensitivity to the Spanish boundary, the Spanish listeners showed a much broader discrimination pattern encompassing both boundaries. In addition, the bilinguals tended to weight the English over the Spanish boundary in both identification and discrimination.

Additional work using only adult English listeners further indicates that *lead* boundaries are less salient than *lag* boundaries and are sometimes not even noticed until specific instruction or training is used to attract attention to them. For example, Pisoni (1977) used a nonspeech continuum consisting of a two-tone complex in which a 500-Hz tone could either *lead* or *lag* a 1500-Hz tone by up to 50 ms. In a two-choice identification test, six of eight listeners showed *lag* boundaries near +20 ms, two listeners showed boundaries near 0 ms, and no listeners showed *lead* boundaries. In a three-choice test, however, after instructions to use three tone categories, listeners learned to identify three categories of tone-onset-time (*lead, simultaneous, lag*), with boundaries near ±20 ms. Pisoni *et al.*, (1982) found similar results with a VOT series (−70 to +70 ms). Given a two-choice label, all listeners showed highly categorical *lag* boundaries near +25 ms. When specifically instructed or trained to use three categories, listeners showed evidence of *lead* boundaries close to −25 ms. In both studies, discrimination tests showed large peaks in the *lag* region, and smaller peaks in the *lead* region.

2. VOT perception in infants and children

Several studies have examined developmental aspects of *lead* versus *lag* perception using synthetic labial CV VOT series.

Three studies used nonnutritive sucking habituation or heart rate deceleration procedures with very young infants (<6 months old). Eimas *et al.* (1971), testing infant English listeners, reported discrimination of a +20/+40-ms *lag* contrast, but no discrimination of a −20/0 *lead* contrast. Lasky *et al.* (1975), testing infant Spanish listeners, reported discrimination of both a *lead* (−20/−60 ms) and a *lag* (+20/+60 ms) contrast. Streeter (1976), testing infant Kikuyu listeners (Kikuyu has a *lead* contrast like Spanish), also reported discrimination of both a *lead* (−30/0 ms) and a *lag* (+10/+40 ms) contrast. This latter author proposed that "previous exposure to the prevoiced-voiced distinction in the language or perhaps exposure to the cues underlying this distinction is a prerequisite for discrimination of the contrast" (p. 40).

Other studies used a headturn procedure to compare *lead* versus *lag* sensitivity in older 6–9 month-old infants. Eilers *et al.* (1979) used a balanced symmetrical design to compare English and Spanish infants briefly tested on two stimulus pairs: −20/+10 (*lead* contrast), and +10/+40 (*lag* contrast). Results showed that the English infants discriminated only the *lag* contrast (92% correct), and did not reliably discriminate the *lead* contrast (46%). However, the Spanish infants discriminated fairly well both the *lead* (80%) and *lag* (88%) contrasts. To account for their results, the authors proposed that "the English boundary has special acoustic properties to which mammalian ears are particularly sensitive," whereas "linguistic listening experience may be a necessary prerequisite to the acquisition of lead VOT contrasts in infants" (Eilers *et al.*, 1979, p. 17).

Aslin *et al.* (1981) also used a headturn procedure to examine more rigorously the sensory capacities of infant (and adult) English listeners for *lead* versus *lag* stimuli. They found that the infants (and adults) could learn to discriminate both a +70 or −70 ms target stimulus from a 0 ms standard. However, additional training and more precise measurements of differential sensitivity using a tracking procedure indicated lower difference limens (better sensitivity) in the *lag* range.

Finally, Streeter and Landauer (1976) used an ABX procedure to measure discrimination in 5–10 year-old children whose native language was Kikuyu (see above), but who were also learning English in school as a second language. When tested with stimulus pairs ranging from −30 to +80 ms, their discrimination of the non-native English *lag* contrast rapidly improved and eventually surpassed that of their native *lead* contrast.

3. VOT perception in animals

Kuhl and colleagues (Kuhl and Miller, 1975; Kuhl and Miller, 1978; Kuhl, 1981; Kuhl and Padden, 1982) conducted a series of very detailed studies to examine VOT perception in chinchillas and monkeys. They used both identification and discrimination tests and various labial, alveolar, and velar series from 0 to +70 ms. In all experiments, the animals showed highly categorical boundaries exactly like those of English listeners, thereby demonstrating that the *lag* boundary is indeed very salient to animals. However, since the *lead* range was never presented, no comparison can be made between *lag* and *lead* sensitivity for the animals.

On the other hand, Waters and Wilson (1976) tested monkey identification with an extended range of VOT (−140 to +140 ms). When presented with various different stimulus ranges, the monkeys exhibited variable boundaries ranging from 11 to 65 ms, and did not perceive the English labial boundary at +25 ms as dominant. However, all monkey boundaries were located on the *lag* portion of the continuum, indicating a preference for *lag* over *lead* boundaries.

Sinnott and Adams (1987) directly compared monkey and adult human English listeners in differential sensitivity along both the *lead* (0 to −70 ms) and *lag* (0 to +70 ms) ranges. Results were very similar to the infant/adult results of Aslin *et al.* (1981): All listeners learned to discriminate both +70 and −70 ms targets from a 0 ms standard, and additional training revealed lower difference limens in the *lag* range.

B. Specific aims of the present study

To summarize, the human research to date has found that the *lead* range of a synthetic labial VOT series, which contains the Spanish boundary, is less salient than the *lag* range, which contains the English boundary. Whereas every listener (whether English, non-English, adult, child, or infant) seems to easily perceive the *lag* boundary with essentially no experience, instruction, or training, not everyone easily perceives *lead* boundaries. Either experience with a *lead* language (e.g., Lasky *et al.*, 1975; Streeter, 1976; Streeter and Landauer, 1976; Williams, 1977a; Eilers *et al.*, 1979) or intensive laboratory training (Aslin *et al.*, 1981; Pisoni *et al.*, 1982; Sinnott and Adams, 1987) appears necessary to enhance perception of less salient *lead* boundaries.

However, the existing animal data are not so clear-cut and raise some questions as to the relative salience of the English versus other possible boundaries on a VOT series. Consider two studies that appear rather contradictory. The Kuhl and Miller (1975, 1978) studies show that animals clearly exhibit the English boundary when presented with only *lag* stimuli. But the Waters and Wilson (1976) study suggests that, when animals are presented with both *lead* and *lag* stimuli, they do not exclusively choose the English boundary. In fact, their behavior may resemble that of various non-English listeners who typically exhibit more variable boundaries [e.g., Caramazza *et al.*, 1973 (French listeners); Keating *et al.*, 1981 (Polish listeners)].

Therefore, our specific aim is to obtain VOT identification data from monkeys, using a combined *lead* and *lag* VOT series, in order to more precisely compare their *lead* and *lag* sensitivity. Two different procedures are used as follows:

1. Boundary "Testing" Procedure

Our first procedure is modeled after the classic two-choice identification test used with human listeners of 2-VOT-category languages (Lisker and Abramson, 1970). Monkeys are first trained to respond differentially to two end points of a VOT series containing both a *lead* and a *lag* range. Next, intermediate stimuli are presented for generalization. The question of interest is whether monkeys choose to place their boundary at the English boundary (as in Kuhl and Miller, 1975), or rather at some other non-English boundary (as in Waters and Wilson, 1976).

2. Boundary "Training" Procedure

In contrast, our second procedure explicitly trains monkeys *via response feedback* to correctly identify VOT stimuli that cross the English *lag* or Spanish *lead* boundaries. Our stimuli are modeled after those used by Eilers *et al.* (1979) in their study of English versus Spanish infant VOT perception. The question of interest is whether monkeys resemble more the "asymmetrical" English infants or the "symmetrical" Spanish infants (or neither) in their perception of the English and Spanish contrasts.

C. Theoretical value of the present study

Monkey perception of VOT in both procedures is then directly compared with that of adult human English and Spanish listeners tested similarly. What is the theoretical value of this three-way design using both a cross language and a comparative comparison within the same study? First, we assume that the monkey gives us a purely psychoacoustic, linguistically unbiased, "baseline" interpretation of the *lead* and *lag* stimuli. Second, we assume that the monkey gives us a reasonably good model of what the prelinguistic human infant perceives at birth. Then, by comparing the monkey's "baseline" performance with that of both native and non-native adult human listeners, we propose to estimate the degree of *perceptual loss* operating in the native listeners or *perceptual learning* operating in the non-native listeners.

We use the Japanese macaque as our "baseline" candidate to model the human prelinguistic infant for two reasons: First, extensive psychoacoustic studies of these monkeys have shown only minor quantitative, as opposed to qualitative, differences with humans with regard to frequency, intensity, and time perception in complex sounds (Sinnott, 1994). Thus their auditory system appears to be basically that of a generalized, as opposed to specialized, primate. Second, extensive studies of these monkeys' vocalizations have not revealed any type of voice-onset-time cues, or other consonantal cues, in their conspecific calls (Green, 1975). Thus they should be basically unbiased in their perception of the human speech feature of VOT.

For purposes of the present study, we therefore propose the following operational definitions of *perceptual loss* and *perceptual learning* in a human linguistic perceiver: *Perceptual loss* in non-native human listeners occurs when the sensitivity of unbiased (monkey) listeners for a particular speech contrast better matches that of native rather than non-native human listeners. In this case, the monkey and native listeners represent the primordial "baseline" perceptual state, and the non-native listeners go on to lose sensitivity. On the other hand, *perceptual learning* in native human listeners occurs when the sensitivity of unbiased (monkey) listeners for a particular speech contrast better matches that of non-native rather than native human listeners. In this case, the monkey and non-native listeners represent the primordial "baseline" perceptual state, and the native listeners go on to develop additional sensitivity.

Now consider specifically the *lead* contrast used in Spanish, which is not easily perceived by English listeners. If non-native English perception of the *lead* boundary involves *perceptual loss*, then monkey perception should be more similar to Spanish than English perception, with monkey and

TABLE I. List of monkey (Monk), American-English (AE), Spanish (SPE=had English 50% boundary, SP=had non-English 50% boundary) listeners. Amount of English (Ex) and Spanish (SPx) exposure is listed in years. For Monk/SPE/SP listeners, age of first exposure to English (FEE) is listed. "na" =not applicable due to being a native listener. B50%=crossover boundary (ms) from the *boundary testing* procedure. For listeners with numerous boundary crossings, the first boundary moving from *lead* to *lag* along the VOT series is listed. PC/lead and PC/lag=PC scores for first five boundary training sessions on *lead* and *lag* stimuli.

Group	Name	Sex	Age	Ex	SPx	FEE	B50%	PC/lead	PC/lag
					A				
Monk	Bongo	M	3	1	0	2	24	91	98
Monk	Dart	M	15	12	0	3	−36	85	98
Monk	Harry	M	15	12	0	3	22	89	92
Monk	Jocko	M	3	1	0	2	23	90	85
Mean			9.0	6.5	0.0	2.5	8.3	88.8	93.3
SD			6.9	6.4	0.0	0.6	29.5	2.6	6.2
					B				
AE	CLG	M	23	na	3	na	25	84	99
AE	JAQ	F	21	na	4	na	27	91	99
AE	KIM	F	29	na	1	na	23	99	100
AE	KIT	F	37	na	1	na	25	60	100
AE	KNT	F	20	na	4	na	27	48	100
AE	SAR	F	41	na	2	na	19	69	100
AE	SHE	F	43	na	0	na	30	78	100
Mean			30.6		2.1		25.1	75.6	99.7
SD			9.7		1.6		3.5	17.8	0.5
					C				
SPE	ANA	F	19	9	na	10	25	88	96
SPE	FER	M	21	5	na	16	27	89	97
SPE	MAG	F	21	11	na	10	23	96	99
SPE	PAT	F	22	10	na	12	25	81	95
Mean			20.8	8.8		12.0	25.0	88.5	96.8
SD			1.3	2.6		2.8	1.6	6.1	1.7
					D				
SP	CAR	M	19	1	na	18	15	97	97
SP	IGA	M	21	3	na	18	−13	98	91
SP	LUI	F	28	2	na	26	11	96	94
SP	MVC	F	18	13	na	5	−11	82	98
SP	PED	M	18	5	na	13	−10	93	86
SP	ROD	M	27	5	na	22	−14	87	93
Mean			21.8	4.8		17.0	**−3.7**	92.2	93.2
SD			4.5	4.3		7.3	13.0	6.4	4.4

Spanish listeners showing enhanced sensitivity relative to English listeners. Specifically, in our *boundary testing* procedure, monkey identification functions should better match Spanish rather than English functions. Also, in our *boundary training* procedure, monkey performance should resemble more the symmetrical pattern of the Spanish infants, rather than the asymmetrical pattern of the English infants (see Eilers *et al.*, 1979).

On the other hand, if native Spanish perception of the *lead* boundary involves *perceptual learning*, then monkey perception should be more similar to English than Spanish perception, with monkey and English listeners showing reduced sensitivity relative to Spanish listeners. Specifically, in our *boundary testing* procedure, monkey identification functions should better match English rather than Spanish functions. Also, in our *boundary training* procedure, monkey performance should resemble more the asymmetrical discrimination pattern of the English infants, rather than the symmetrical pattern of the Spanish infants (see Eilers *et al.*, 1979).

II. METHOD

A. Subjects

The monkey listeners were four male Japanese macaques (*Macaca fuscata*). Dart and Harry were 15-year-old adults with 11 years of previous experience on various speech perception experiments, none involving VOT. Bongo and Jocko were 3-year-old juveniles who were just beginning their careers in speech perception and had only just completed absolute threshold testing and initial shaping on the left-right procedure using the vowels /u/ and /i/. All the monkeys had frequent informal exposure to English in the lab from interactions with humans, but had no exposure to Spanish (see Table I, Part A). All experimental procedures for animals were approved by the Animal Care and Use Committee of the University of South Alabama.

The human listeners were students or staff members of the University of South Alabama. Of the American-English (AE) group ($n=7$), several had been exposed briefly to Spanish in high school or college, but none considered themselves

fluent (see Table I, Part B). The Spanish group ($n=10$) all came from South America, had Spanish as their first language, and spoke English with noticeable Spanish accents. All except one (MVC) had received their first exposure to English after the age of 9 years (see Table I, Parts C and D). Several were recruited from the University's English as a Second Language (ESL) Institute and were not very fluent in English. Human subjects were paid $10 for each daily 15-min session.

B. Stimuli

The VOT stimuli were modeled after those developed by Pisoni *et al.* (1982). They were synthesized (10 kHz rate) using the Klatt (1980) synthesizer contained in the Computerized Speech Research Environment (AVAAZ Innovations, Ontario Canada) and consisted of 14 synthetic labial stop consonant-vowel (CV) syllables ranging from −60 to +70 ms in 10-ms steps. All stimuli had a 200-ms steady-state vowel portion with formants appropriate for /a/ (F1=730 Hz, F2 =1090 Hz, F3=2240 Hz). F0 was flat at 125 Hz until the last 100 ms of the vowel when it decreased to 100 Hz. The initial transitions into the vowel were 40 ms long with starting frequencies as follows: F1=500 Hz; F2=800 Hz; F3 =1900 Hz. All stimuli contained a 10-ms release burst that was generated using the frication source (AF), which has a broadband (5 kHz) flat spectrum, and the bypass channel (AB). Voicing *lead* was simulated by passing a combination of the sinusoidal voicing source and normal voicing through F1 (180 Hz; BW1=150 Hz). Aspiration in the *lag* region was generated using the aspiration source (AH) and a F1 bandwidth widened to 300 Hz. All other synthesis parameters were left at the default values.

C. Apparatus

The apparatus has been previously described (Sinnott and Saporita, 2000). Briefly, testing took place in the free-field inside an Industrial Acoustics Corp. (IAC) booth (1.98 m × 1.82 m × 1.52 m) lined with sound-absorbing material. Monkeys sat in a primate chair and responded on an attached lever that could be moved left or right. Humans sat in a normal human chair and responded on a lever mounted on a post in front of them. A cue light above each lever signaled the listener to contact the lever to begin a trial. The A-D converter, interfaces, 8k-Hz filter, and attenuators were all Tucker-Davis-Technologies equipment. Stimulus delivery, experimental contingencies, and response recording were controlled by a Dell computer.

Stimuli were presented at a normal conversational level of 65 dB SPL via a Genesis loudspeaker located approximately 84 cm from the listener's head. The level at 180 Hz (i.e., the region of voicing *lead*) was calibrated (using a pure tone) to be at 55 dB SPL, which is well within the hearing range of both humans and monkeys (Owren *et al.*, 1988).

D. Basic go-left/go-right identification procedure

For both the *boundary testing* and *boundary training* procedures (see Secs. I B 1 and I B 2 above; II D 1 and II D 2 below), a *go-left/go-right*, 2AFC identification procedure was used (Sinnott and Saporita, 2000). A trial began with the flashing cuelight, which signaled the listener to contact the lever. At contact, a variable hold of 1–2 sec occurred, and then a stimulus started to repeat indefinitely at a rate 1 per sec until a *go-left* or *go-right* response was made. Thus a listener could take as much time as desired to listen to each stimulus on each trial. If the lever was pushed in the correct direction, feedback was delivered in the form of a 100-ms, 2-kHz tone pip, and a banana pellet for monkeys. If the lever was pushed in the wrong direction or simply released, a miss resulted and a 5-s timeout, during which a 300-Hz tone was presented. Each miss resulted in activation of a correction procedure, such that missed stimuli were repeated on successive trials until a correct response occurred. Correction procedure trials occurring after the initial miss were not counted in the data analysis; their sole purpose was to prevent the development of *go-left* or *go-right* response biases from developing in any listeners.

1. Boundary "Testing" Procedure

The first step was to train all listeners to reliably identify the two VOT end points (−60 and +70 ms). A *go-left* response was required for the extreme −60 ms *lead* stimulus, and a *go-right* response was required for the extreme +70 ms *lag* stimulus.

Monkeys were shaped using an intensity cue (Sinnott and Saporita, 2000). The −60 ms (*go-left*) stimulus was initially reduced in intensity by 20 dB relative to the +70 ms (*go-right*) stimulus, but it was still audible. When performance on a given day averaged $>=85\%$ correct, the level of the −60 ms stimulus was increased by 2 dB for the next daily session. Shaping continued until both stimuli were presented at equal intensity (65 dB SPL), typically requiring 10–20 days.

Humans were introduced to the end point stimuli with no intensity cues. They were simply given verbal instructions as follows: "When you see a flashing light, contact the lever. At contact, the light will stop flashing and after a short time the stimulus will begin to repeat indefinitely. You must move the lever left for 'BA' and right for 'PA.' If you make a correct response, you will hear a high frequency beep. An incorrect response will place you in a 5-s timeout and you will hear a low frequency buzz." Author L.A.P. (Laura A. Powell) gave instructions in English to AE listeners; author J.C. (Jazmin Camchong) who is a native speaker of Spanish from Ecuador, gave instructions in Spanish to Spanish listeners. Then one practice session using the −60 and +70 ms stimuli was presented to familiarize them with the procedure.

After the initial shaping (monkeys) or practice (humans) with the end point stimuli, all subjects were then tested to see how they generalized to identify the intermediate stimuli (−50 through +60 ms) and thus determine the "crossover" boundary at which 50% *go-left* and 50% *go-right* responses occurred. During generalization testing, there was only feedback for responses to the end point stimuli, and for these stimuli the correction procedure was in effect. For the intermediate stimuli, all *go-left* or *go-right* responses were treated as correct, and the correction procedure was not in effect. The end point stimuli occurred on 33% of the trials, and the

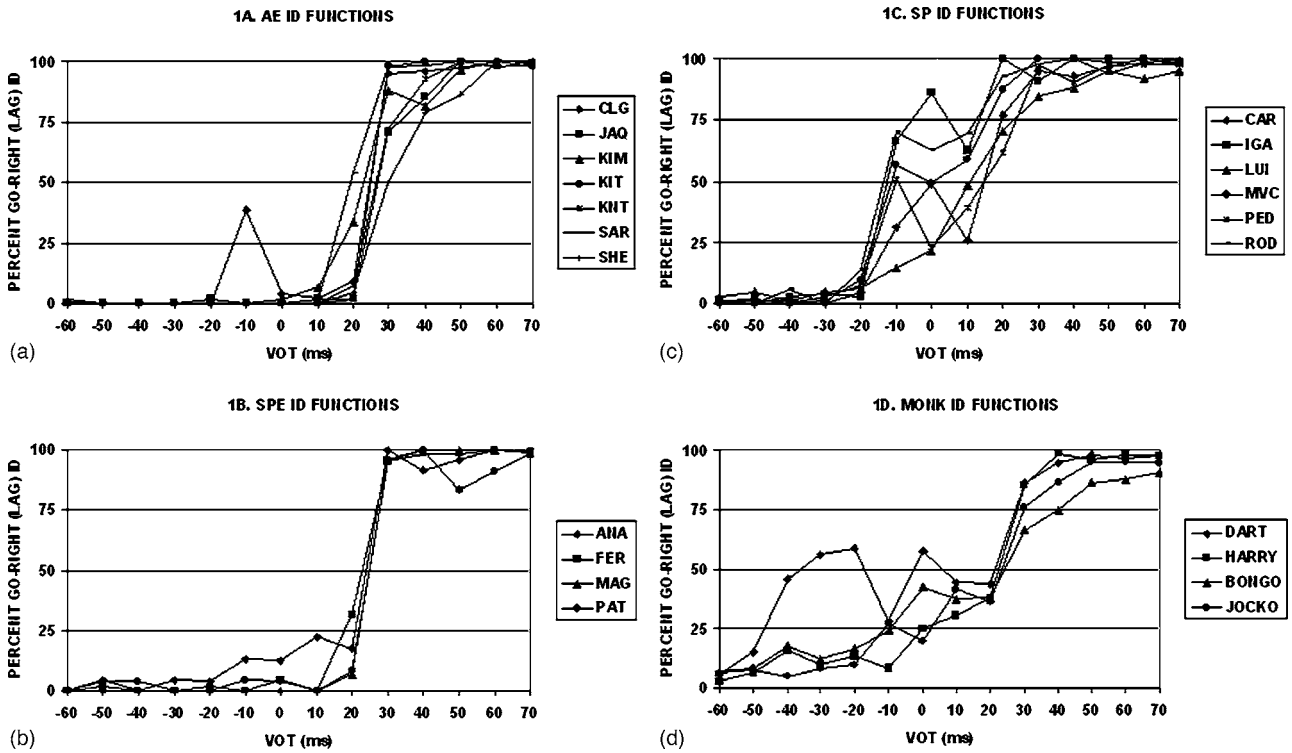


FIG. 1. Individual two-choice identification functions along the VOT series from −60 to +70 ms, as measured during the *boundary testing* procedure. (a) Seven AE listeners. (b) Four SPE listeners with boundaries inside the AE range. (c) Six SP listeners with boundaries outside the AE range. (d) Four monkey listeners.

intermediate stimuli on 67% of the trials. Each daily session consisted of 120 trials, resulting in about 40 end-point stimuli and 80 intermediate stimuli. A total of 10 sessions were obtained, resulting in a grand total of about 50 responses for each intermediate stimulus.

2. Boundary "Training" Procedure

The English stimuli consisted of +10 ms VOT (requiring a *go-left* response), and four different *lag* stimuli: +40, +50, +60, +70 ms VOT (all requiring a *go-right* response). The Spanish stimuli consisted of +10 ms VOT (requiring a *go-left* response), and four different *lead* stimuli: −20, −30, −40, −50 ms VOT (all requiring a *go-right* response). Thus the subject always had to *go-left* to +10 ms VOT, but the *go-right* stimuli varied between crossing the *lead* and *lag* boundaries, and also incorporated varying degrees of proximity to the boundaries. Stimuli requiring *go-left* and *go-right* responses were randomly presented.

Monkeys were initially shaped to perform on both the *lead* and *lag* stimuli using an intensity cue, which initially attenuated the *go-left* (+10 ms) stimulus by −20 dB relative to the *go-right* stimuli. Shaping alternated daily between *lead* and *lag* conditions. Within each condition, the intensity cue was reduced by 2 dB per session when performance averaged $>=85\%$ correct. All monkeys learned to differentiate the *lag* stimuli at equal intensity in the minimum number of shaping sessions (10), while it took them longer to differentiate the *lead* stimuli (10–22 sessions). But note that *lead* shaping involved a confounding variable, since *boundary testing* had previously required subjects to *go-left* to *lead* stimuli, while now *boundary training* required them to *go-right*. For this reason, we could not legitimately analyze "trials to criterion" by monkeys as a measure for comparing *lead* and *lag* processing. Actual data acquisition began on the first session in which the stimuli within each of the *lead* and *lag* conditions reached equal intensity, when monkeys could no longer use the intensity cue to differentiate the stimuli.

For humans, all *lead* and *lag* stimuli were initially introduced at equal intensity level. They were told that daily test sessions would alternate between an English versus Spanish contrast, and were instructed to "listen very carefully to the stimuli," and to "try your best to differentiate them." L.A.P. gave instructions to English listeners in English; J.C. gave instructions to Spanish listeners in Spanish. Actual data acquisition began after one practice session on each of the *lead* and *lag* conditions.

After the initial shaping (monkeys) or practice (humans) sessions, each daily training session consisted of approximately 120 trials. This resulted in about 60 presentations of the single *go-left* stimulus (+10 ms), and about 15 presentations of each of the four *go-right* stimuli (VOT *lead*: −20, −30, −40, −50; VOT *lag*: +40, +50, +60, +70). A total of 30 sessions were obtained for data analysis (15 *lead* alternating with 15 *lag*). Response times to correct responses were also recorded.

III. RESULTS

A. Boundary Testing Procedure

1. Individual data

All individual identification functions are shown in Figs. 1(a)–1(d). All individual boundaries are also listed numeri-

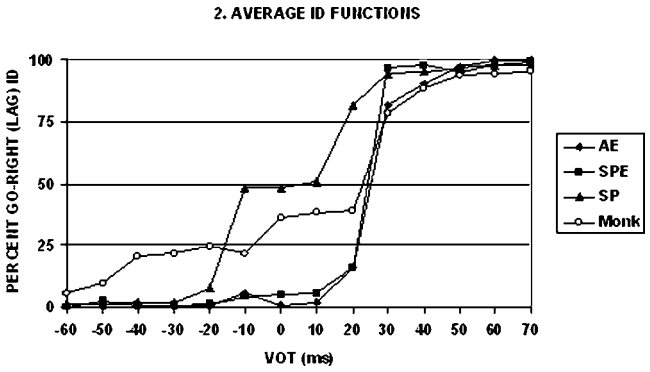


FIG. 2. Averaged human AE, SPE, SP, and monkey identification functions, showing "plateaus" in the SP and monkey functions.

cally in Table I, Column "B50%," along with group means and standard errors.

Figure 1(a) shows the AE functions. Except for one aberrant function at −10 ms, all are highly categorical with 50% crossover boundaries ranging between 19–30 ms.

In contrast to the AE functions, the Spanish functions were not homogeneous. Of the ten Spanish listeners, four (ANA, FER, MAG, PAT) showed categorical English-like perception, and had boundaries at 23–25 ms, values that were within the AE range of 19–30 ms (see Table I). This group is referred to as the Spanish-English (SPE) group, and their functions are shown in Fig. 1(b). The other six showed variable non-English boundaries ranging between −14 and +15 ms, values that were outside of the AE range of 19–30 ms (see Table I). This latter group is referred to simply as the Spanish (SP) group, and their functions are shown in Fig. 1(c).

Figure 1(d) shows the four monkey functions. Dart shows an extremely nonmonotonic function that crosses the 50% point several times, first at −35 ms and last at +20 ms, near the English boundary. While the other monkey functions have a single 50% crossover boundary in the AE range, they are clearly not categorical like the AE and SPE functions.

2. Average data

Figure 2 shows average functions for the four groups (AE, SPE, SP, monkey). The AE and SPE functions both have highly categorical *lag* boundaries near +25 ms VOT, while the monkey and SP functions are more variable and both exhibit "plateaus" near the middle of the series. In addition, the monkey functions show more "lag" responding in the *lead* range of −60 to −20 ms, unlike any human functions. To determine whether these group effects were significant, all percentages were arc-sine transformed and the four listener groups (AE, SPE, SP, monkey) were compared on the 14 levels of VOT (−60 to +70 ms), using a 4×14 analyses of variance (ANOVA). All effects were significant: GROUP ($F=15.056[3,17]$, $p=0.000$); VOT ($F=451.197$ [13,221], $p=0.000$); GROUP × VOT($F=11.853[39,221]$, $p=0.000$).

In further analyses, the monkey functions were compared separately with those of each of the three human groups, again using the entire VOT range. For all comparisons, the effect of VOT was significant and is not further discussed, simply indicating increased "lag" responding as VOT increased from −60 to +70 ms. The GROUP comparisons revealed that the monkey functions were significantly different from the AE ($F=14.893[1,9]$, $p=0.004$) and the SPE ($F=6.597[1,6]$, $p=0.042$), but not from the SP ($F=0.991[1,8]$, $p=0.349$) functions.

3. Group comparisons in the lead versus simultaneous versus lag ranges

Next, the VOT series was partitioned into three ranges, based on the "nonspeech" temporal order thresholds determined by Pisoni (1977). These were: a *lead* range containing four VOTs from −60 to −30 ms, a *simultaneous* range containing five VOTs from −20 to +20 ms, and a *lag* range containing five VOTs from +30 to +70 ms. In all cases, the effect of VOT was significant and is not further discussed. The GROUP comparisons revealed significant differences in the *lead* ($F=9.108[3,17]$, $p=0.001$) and *simultaneous* ($F=26.882[4,12]$, $p=0.000$) ranges, but not in the *lag* ($F=2.057[3,17]$, $p=0.144$) range. Thus the various listener groups differed as to whether to call stimuli "lead" or "lag" in the *lead* and *simultaneous* ranges, but all were similar in calling stimuli "lag" in the *lag* range.

The monkey functions in each VOT range were then compared separately with those of each human group. For all comparisons, the effect of VOT was significant and is not further discussed.

Lead range: Monkey functions differed from those of all human groups (Monk versus AE: $F=13.055[1,9]$, $p=0.006$; Monk versus SPE: $F=5.716[1,6]$, $p=0.054$; Monk versus SP: $F=8.310[1,8]$, $p=0.020$). Thus monkeys sometimes called these stimuli "lag," whereas humans always called them "lead."

Simultaneous range: Monkey functions differed from AE ($F=35.530[1,9]$, $p=0.000$) and SPE ($F=18.538[1,6]$, $p=0.005$), but not from SP functions ($F=3.549[1,8]$, $p=0.096$). Both monkey and SP listeners called more stimuli "lag" in this range, compared to AE and SPE listeners.

Lag range: As indicated above, there were no significant differences between monkey functions and those of any human group (Monk versus AE: $F=0.962[1,9]$, $p=0.352$; Monk versus SPE: $F=4.571[1,6]$, $p=0.076$; Monk versus SP: $F=4.776[1,8]$, $p=0.060$). All groups were similar in calling the stimuli "lag" in this range.

B. Boundary Training Procedure

1. Early boundary training identification performance

Figure 3(a) shows a comparison of mean PC (percent correct) identification for the first ten training sessions (5 *lead*, 5 *lag*) for AE, SPE, SP, and monkey listeners. All individual data are also listed in Table I, Columns "PC lead" and "PC lag," along with group means and standard errors. The PC data were arcsine-transformed and analyzed using a 4(GROUP) × 2(LEAD/LAG) ANOVA.

The main effect of GROUP was not significant ($F=1.0[3,17]$, $p=0.417$), indicating that all groups including monkeys were performing at similar overall levels of accu-

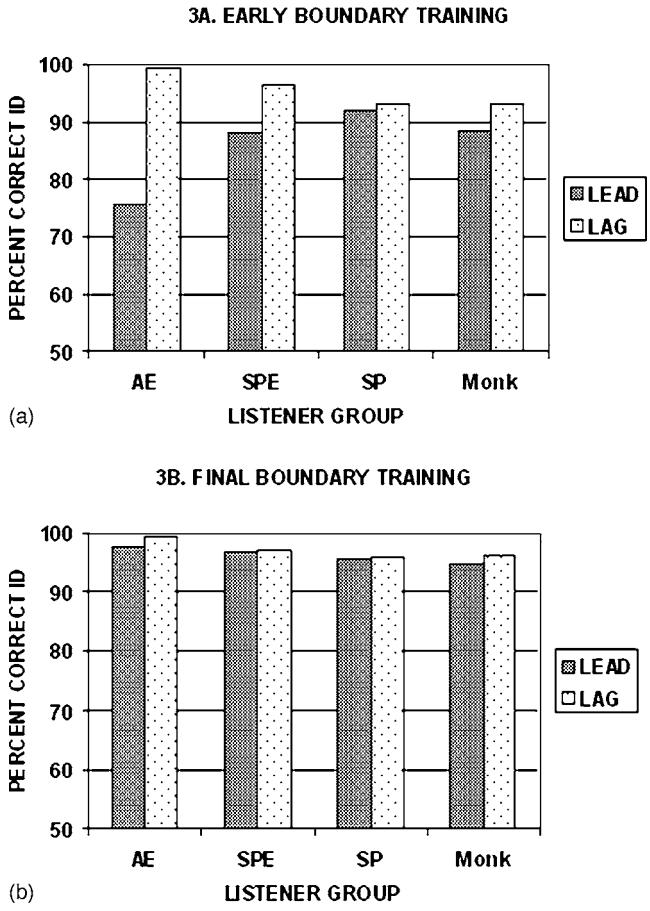


FIG. 3. (a) Percent correct identification for the four listener groups (AE, SPE, SP, Monk) comparing the *lead* versus *lag* contrasts during the early stage of *boundary training*. (b) Data obtained during the final stage of *boundary training*.

racy in this early stage of training. The effect of LEAD versus LAG was significant ($F=11.694[1,17]$, $p=0.003$), indicating overall lower (PC) scores with the *lead* stimuli. The interaction was also significant ($F=4.411[3,17]$, $p=0.018$).

To determine the source of the interaction, the effect of LEAD versus LAG was then compared by a T-test within each listener group. For AE and SPE identification, the effect was significant (AE: $t=-3.535$, $[6]p=0.012$; SPE: $t=-3.667[3]$, $p=0.035$). For SP and monkey identification, the effect was nonsignificant (SP: $t=-0.278[5]$, $p=0.792$; MNK: $t=-1192[3]$, $p=0.319$). Thus the AE and SPE listeners had significantly asymmetrical processing of the *lead* versus *lag* stimuli, with reduced sensitivity to the *lead* stimuli. In contrast, the SP and monkey listeners did not show significantly asymmetrical processing of the same stimuli.

2. Final boundary training identification

The second measure examined was mean PC identification for the last ten training sessions (5 *lead*, 5 *lag*). One AE listener (KNT) dropped out of the experiment after performing at chance level on the *lead* stimuli during early training. So these data are based on the remaining six AE listeners. Figure 3(b) shows a comparison of mean PC identification for the last ten training sessions (5 *lead*, 5 *lag*) for AE, SPE, SP, and monkey listeners. At the end of training, all groups including monkeys showed comparable asymptotic PC identification ($>=95\%$) and there were no significant effects: GROUP ($F=2.828[3,16]$, $p=0.072$); LEAD versus LAG ($F=2.563[1,16]$, $p=0.129$); Interaction: ($F=0.607[3,16]$, $p=0.620$).

3. RT data at PC asymptote

The next measure examined was mean response times (RT) for the final ten training sessions, since PC scores all were at asymptotic levels (see Sec. III B 2 above). RTs were converted to log RTs for statistical analysis, to equate proportional changes in RT for listeners with overall slower or faster response times. There was no effect of GROUP ($F=1.253[3,16]$, $p=0.324$); a significant effect of LEAD/LAG ($F=10.629[1,16]$, $p=0.005$); and no interaction ($F=0.361[3,16]$, $p=0.782$). Thus there were no listener group differences in speed of responding to *lead* versus *lag* stimuli and all groups were similar in having longer RTs to the *lead* stimuli (Mean RTs for lead=647 ms; Mean RTs for lag =602 ms).

4. Detailed analysis of the LEAD and LAG stimuli

The last measure examined involved a detailed analysis of the *lead* and *lag* stimuli during the final ten boundary training sessions. Recall that the *go-left* stimulus was always +10 ms VOT, and the *go-right* stimuli were −20, −30, −40, and −50 ms (*lead* contrast), or +40, +50, +60, and +70 ms (*lag* contrast). These values reflect increasing distances from the Spanish *lead* boundary at −10 ms, or from the English *lag* boundary at +25 ms. The purpose of these analyses was to determine if any listener groups would exhibit more/less accurate identification for stimuli that were further/nearer to the *lead* or *lag* phoneme boundaries. In each condition, possible effects were analyzed with a 4(GROUP)×5(STIM) ANOVA.

PC identification by the four listener groups for the five stimuli in the *lead* condition is shown in Fig. 4(a). All groups performed at >90% correct for all stimuli. No effects were significant: GROUP ($F=1.314[3,16]$, $p=0.304$); (STIM) ($F=1.289[4,64]$, $p=0.284$); interaction ($F=1.061[12,64]$, $p=0.407$). Thus all listener groups had similar levels of accuracy and patterns of detailed responding to the *lead* stimuli in the final stages of boundary training.

PC identification by the four listener groups for the five stimuli in the *lag* condition is shown in Fig. 4(b). All groups except the SP listeners performed at >90% correct for all stimuli. Here all effects were significant: GROUP ($F=7.613[3,16]$, $p=0.002$), reflecting higher PC identification for AE and SPE than SP and monkey listeners; STIM ($F=10.149[4,64]$, $p=0.000$), reflecting lower PC identification for the +40 ms stimulus; Interaction: ($F=3.127[12,64]$, $p=0.002$).

To determine the source of the interaction, PC identification for the five stimuli (+10, +40, +50, +60, +70 ms) was then compared within each listener group using a series of one-way ANOVAs. There was no effect of STIM on AE ($F=2.337[4,20]$, $p=0.086$), SPE ($F=1.878[4,12]$, $p=0.179$, and monkey ($F=1.686[4,12]$, $p=0.218$) identifica-

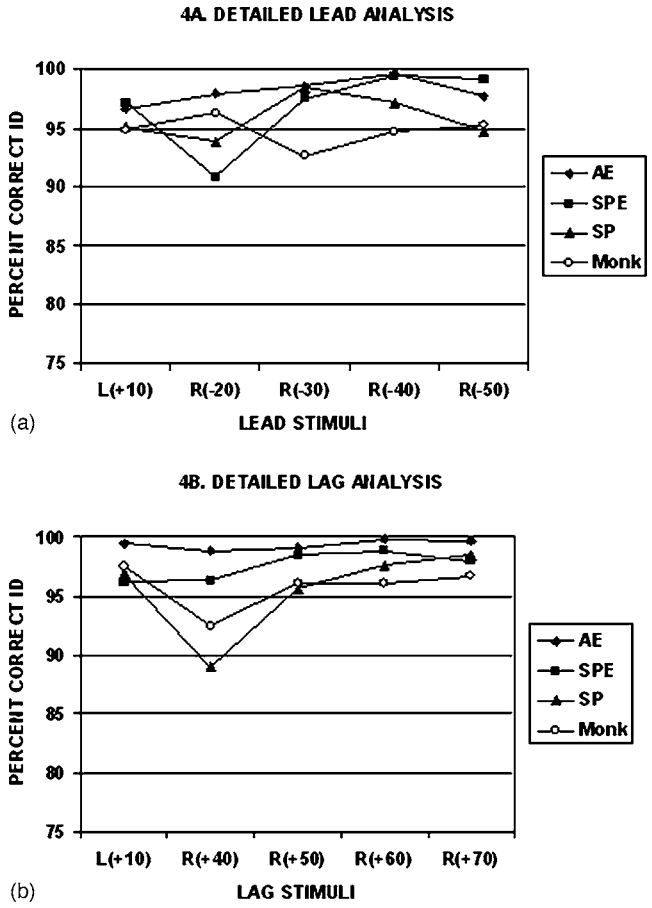


FIG. 4. (a) Detailed analysis of the percent correct identification comparing the individual *go-left* [L] and *go-right* [R] LEAD stimuli during the final stage of *boundary training*. (b) Similar analysis comparing the individual *go-left* [L] and *go-right* [R] LAG stimuli.

tion. However, the STIM effect was significant for SP identification ($F=10.035[4,20]p=0.000$), indicating lower PC identification for the +40 ms stimulus closest to the *lag* boundary.

IV. DISCUSSION

A. Boundary testing: Which VOT boundaries do monkeys choose?

As proposed above, we view monkeys as linguistically unbiased listeners who would yield a purely psychoacoustic interpretation of the possible *lead* or *lag* boundaries along a synthetic VOT series. The question of interest in our *boundary testing* procedure was whether monkeys would choose to place their boundaries at the English boundary (as in Kuhl and Miller, 1975), or rather at some other non-English boundary (as in Waters and Wilson, 1976).

Our results indicated that three of four monkeys chose the English boundary, since their 50% boundaries were within the AE boundary range (19–30 ms). But one monkey Dart showed several boundaries, with some in the *lead* range. In addition, no monkey functions were truly categorical like those of Kuhl and Miller (1975). Recall that Kuhl and Miller presented only the *lag* range to their chinchillas and found clear categorical functions that matched those of English listeners. Looking at our Fig. 1(d), most likely our monkeys would have to if presented with only the *lag* range. But when presented with an extended series, our monkeys showed variable non-English functions that better matched the SP than AE functions.

The most interesting comparison occurs in the *simultaneous* range between approximately −20 and +20 ms. (see Fig. 2). Here both the monkey and SP functions have "plateaus" indicating sensitivity to three types of VOT cues: *lead, simultaneous*, and *lag*. In contrast, the AE and SPE functions reveal sensitivity to only two types of VOT cues, with a single clear boundary at +25 ms, cleanly separating the combined *lead/simultaneous* range from the *lag* range.

In fact, during the course of testing, several SP listeners "complained" to author J.C. that they did in fact perceive three VOT categories in the stimuli and were unsure what to do with the middle category in a two-choice procedure. In any case, the variable identification functions obtained from our monkey and SP listeners are not unusual and resemble those found in other studies of non-English VOT perception (e.g., Caramazza *et al.*, 1973, Keating *et al.*, 1981).

Also note one important way in which the monkey functions were different from all human functions. In the *lead* range, the monkeys had a higher rate of "lag" responding compared to all human groups. This behavior could be an artifact of our shaping procedure used to train monkey responses to the end points of the series, since the *go-left* −60 ms *lead* stimulus was initially reduced by −20 dB relative to the *go-right* +70 ms *lag* stimulus. But an opposing argument is that, when all stimuli were presented at the same level during generalization testing, the monkeys were equally and symmetrically accurate in responding to the two end point stimuli.

Finally, note one important way in which the monkey functions were similar to all human functions. In the *lag* range, all listeners were similar in labeling the stimuli as "lag." This pattern seems to indicate a highly salient universal English *lag* category based on some sort of general mammalian auditory mechanism.

B. Boundary training: Are monkeys more like English or Spanish infants?

The question of interest in our *boundary training* procedure was whether our monkeys would resemble more the "asymmetrical" English infants or the "symmetrical" Spanish infants of Eilers *et al.* (1979) in their perception of English and Spanish VOT contrasts.

First consider the present data comparing monkeys with AE and SP adult human listeners. In the early stages of training, before identification performance reached asymptotic levels, monkey and SP identification accuracy was more symmetrical for the *lead* versus *lag* stimuli, compared to AE identification [see Fig. 3(a)]. These data should be viewed with some caution, because the monkeys (unlike humans) were shaped to differentiate the *lead* and *lag* stimuli using intensity cues, and this preliminary shaping could well have influenced their more symmetrical discrimination performance. However, note that these early training results seem to verify the initial (*boundary testing*) monkey identification functions that reveal spontaneously detected differences be-

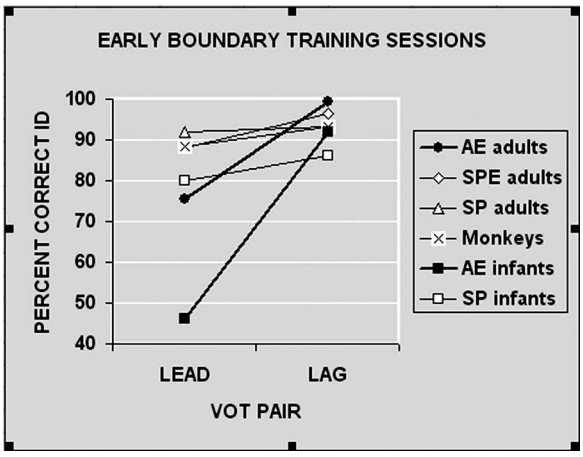


FIG. 5. Comparison of the present early *boundary training* data [replotted from Fig. 3(a)] with the human infant data of Eilers *et al.* (1979). Listener groups for the present (adult) study: AE=American-English; SPE=Spanish with English ID function; SP=Spanish with non-English ID function; Monk=monkey. Listener groups for the Eilers (infant) study: AE infant=English-learning infant; SP infant=Spanish-learning infant.

tween *lead* and *simultaneous* stimuli not shown in the AE functions (see Fig. 2).

Now consider the present early *boundary training* data [replotted from Fig. 3(a)] in comparison with the English and Spanish human infant data of Eilers *et al.* (1979). All these data are shown together in Fig. 5. A slope is plotted for each listener group comparing relative identification or discrimination accuracy on *lead* versus *lag* stimuli. The two groups with the most asymmetrical slopes are the Eilers' infant English listeners and our AE adults. In addition, our monkeys have an almost identical slope to Eilers' infant Spanish listeners. This cross-study comparison suggests that, at least in the early stages of training, not only does monkey perception of VOT provide a better match to Spanish than English adult VOT perception, but it also provides a better match to Spanish than English infant VOT perception.

C. Using monkeys to uncover perceptual loss in human speech perception

As proposed above, the main theoretical goal of this study is to gather evidence for *perceptual loss versus perceptual learning* models in human speech perception development. Recall that we consider *perceptual loss* in a non-native human listener to occur when monkey "baseline" sensitivity provides a better match to native than non-native sensitivity. In contrast, we consider *perceptual learning* in a native human listener to occur when monkey "baseline" sensitivity provides a better match to non-native than native sensitivity.

Our first set of results from *boundary testing* (see Fig. 2) suggest that monkey sensitivity in the *lead/simultaneous* region of the VOT series provides a better match to native SP sensitivity than to non-native AE sensitivity. Specifically, our monkey and SP functions suggest sensitivity to three types of acoustic cues in the VOT series (*lead versus simultaneous versus lag*), while the AE functions suggest sensitivity to only *two* types (*lead/simultaneous versus lag*). Thus these data provide evidence for *perceptual loss* operating in the AE listeners for the *lead versus simultaneous* contrast.

Our second set of results from early *boundary training* [see Fig. 3(a)] also suggest that monkey sensitivity provides a better match to native SP sensitivity than to non-native AE sensitivity. Specifically, our monkey and SP performance was more symmetrical for the *lead* and *lag* boundaries, while our AE performance was more asymmetrical. Thus these data provide additional evidence for *perceptual loss* operating in our AE listeners for the *lead versus simultaneous contrast.*

Third, our final evidence of *perceptual loss* is concerned with SP perception of the English *lag* contrast. In the detailed analysis of the *lag* contrast, monkey performance better matched native AE than non-native SP performance. Here, the SP listeners were the only ones to show significantly reduced accuracy to the +40 ms stimulus closest to the English boundary [see Fig. 4(b)]. This effect suggests that SP identification of the *lag* category was less accurate and less categorical compared to AE and monkey identification, suggesting *perceptual loss* in the SP listeners for this particularly subtle aspect of *lag* perception.

D. The case of the human SPE listeners

In our statistical analyses, SPE identification provided a better match to AE patterns, rather than to those of their fellow SP listeners. Specifically, in *boundary testing*, they chose a non-native English over their native Spanish boundary, and in the early stages of *boundary training*, they had asymmetrical identification of the *lead* versus *lag* boundaries, reflecting reduced sensitivity to their native *lead* boundary. If Spanish was their first, native language, and they had Spanish accents in production, then why would our SPE listeners not have accompanying Spanish perception in the present experiment?

First note that the SPE listener group had an earlier mean age of first exposure to English [12 years old) than the SP listener group [17 years old; see column "first exposure to English" (FEE) in Table I]. Thus there is some basis for differentiating the SPE/SP results by date of first exposure to English. But also note that while SP MVC has the earliest age of exposure to English (5 years old), she clearly has a negative non-English boundary (−11 ms VOT). Because of inconsistencies like this one that make it difficult to exactly correlate FEE with SP or SPE perception, we feel it is beyond our level of expertise to investigate further why our native Spanish listeners divided into SP and SPE groups.

Also note that the present SPE results seem to be similar to those of Streeter and Landauer (1976), who reported that native Kikuyu children weight the English *lag* boundary over their native *lead* boundary, at least after becoming exposed to English in school. We propose two possible hypotheses to explain our strikingly similar results:

First, the *lead* VOT contrast is probably much less salient acoustically than the *lag* contrast. For example, the low-frequency voice bar that cues voicing *lead* is in a region of diminished primate auditory sensitivity (<200 Hz), while the aspiration noise that cues voicing *lag* is in a region of enhanced auditory sensitivity (500–2000 Hz) (Owren *et al.*, 1988). So perhaps a less salient native contrast might come

to be dominated by a competing more salient non-native contrast during second language learning.

Second, the present VOT series (as well as all those used previously) might not accurately capture all the relevant perceptual cues contained in the Spanish *lead* contrast. For example, Williams (1977b) presents perceptual data showing that Spanish /pa/ cannot be changed into /ba/ by simply adding a voice bar. Yet this has always been the method of synthesizing a prevoiced /ba/ in all VOT series synthesized so far (ours included). So perhaps our SP listeners were able to "assimilate" (e.g., Best, 1994) these imperfect examples of the Spanish contrast into their native categories, while our SPE listeners were not.

E. Need for data from monolingual Spanish adult listeners

In the present study, all our native Spanish listeners had received some exposure to English, so they were not truly monolingual Spanish listeners. How would totally monolingual adult Spanish listeners perform in the present procedures? Regarding *boundary testing*, the data from Williams (1977a) seem to indicate clear categorical identification functions with a mean boundary at −10 ms VOT, clearly different from our monkey (or SP) data. Thus it is likely that further research would show that monkeys would not provide a good match to either monolingual adult English or Spanish listeners. This result in turn would suggest that both mechanisms of *perceptual loss* (operating in non-native English listeners) as well as *perceptual learning* (operating in native Spanish listeners) contribute to development of the Spanish *lead* contrast. Note however, with regard to early *boundary testing*, that the present data do strongly suggest that monkeys provide a better match to monolingual Spanish infant data than to monolingual English infant data (see Fig. 5).

V. CONCLUSION

The present results suggest that monkeys can be used to explore *perceptual loss* versus *perceptual learning* in human speech perception. However, the most important limitations of the present study are first, since VOT series as currently synthesized may not contain all the cues to Spanish voicing, it would be valuable to replicate the present *boundary training* data using real speech stimuli. Second, monolingual Spanish adult listeners should be tested to further investigate the role of *perceptual learning* in the case of the *lead* contrast. Third, there are obvious limitations in using a two-choice procedure to categorize a VOT series with three phonetic categories of *lead*, *simultaneous* and *lag*. But unfortunately, three-choice procedures do not yet exist for animals, so three-category experiments comparing humans and animals are still very much in the future.

Despite these limitations, the present results are still informative. First, they can be compared with much previous research using similar VOT series, as in our Fig. 5. Second, while a two-choice procedure in *boundary testing* may lead to variable boundaries for non-English (e.g., certain Spanish and monkey) listeners, the fact remains that English listeners have very stable, categorical boundaries with the exact same procedure. This tells us something: Monkeys do not automatically pick up English phonology via passive exposure to English, becoming inattentive to voicing *lead*. (Consider that our two oldest monkeys Dart and Harry have been listening to English in the lab for over 15 years.) So monkeys seem to remain unbiased linguistic observers, attending to both *lead* and *lag*, and giving a more-or-less generalized psychoacoustic interpretation of speech stimuli. Thus monkeys should prove to be of considerable value in comparisons with human infants in the process of language learning.

ACKNOWLEDGMENTS

This research was supported by R01 DC00541-14 to J.M.S. These data were based on a Master's thesis by L.A.P. in the Psychology Department at the University of South Alabama. The authors thank Kelly Mosteller and Casey Gilmore for assistance in testing monkeys.

Aslin, R., Pisoni, D., Hennessy, B., and Perey, A. (**1981**). "Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience," Child Dev. **52**, 1135–1145.

Best, C. (**1994**). "The emergence of native-language phonological influences in infants: A perceptual assimilation model ," in *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words*, edited by J. Goodman and H. Nusbaum, pp. 167–223 (MIT Press, Cambridge, MA).

Caramazza, A., Yeni-Komshian, G., Zurif, E., and Carbone, E. (**1973**). "The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals," J. Acoust. Soc. Am. **54**, 421–428.

Eilers, R., Gavin, W., and Wilson, W. (**1979**). "Linguistic experience and phonetic perception in infancy: A crosslinguistic study," Child Dev. **50**, 14–18.

Eimas, P., Siqueland, E., Josczyk, P., and Vigorito, J. (**1971**). "Speech perception in infants," Science **171**, 303–306.

Green, S. (**1975**). "Variation of vocal pattern with social situation in the Japanese monkey (*Macaca fuscata*): A field study," in *Primate Behavior, Vol. 4: Developments in Field and Laboratory Research*, edited by L. A. Rosenblum, pp. 1–102 (Academic Press, NY).

Keating, P., Mikos, M., and Ganong, W. (**1981**). "A cross-language study of range of voice onset time in the perception of initial stop voicing," J. Acoust. Soc. Am. **70**, 1261–1271.

Kuhl, P. (**1981**). "Discrimination of speech by non-human animals: Basic auditory sensitivities conducive to the perception of speech-sound categories," J. Acoust. Soc. Am. **70**, 340–348.

Kuhl, P., and Miller, J. D. (**1975**). "Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants," Science **190**, 69–72.

Kuhl, P., and Miller, J. D. (**1978**). "Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli," J. Acoust. Soc. Am. **63**, 905–917.

Kuhl, P., and Padden, D. (**1982**). "Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques," Percept. Psychophys. **32**, 542–550.

Lasky, R., Syrdal-Lasky, A., and Klein, R. (**1975**). "VOT discrimination by four to six and a half month old infants from Spanish environments," J. Exp. Child Psychol. **20**, 213–225.

Lisker, L., and Abramson, A. (**1964**). "A cross-language study of voicing in initial stops: Acoustical measurements," Word **20**, 384–422.

Lisker, L., and Abramson, A. (**1970**). "The voicing dimension: Some experiments in comparative phonetics," in *Proceedings of the 6th International Congress of Phonetic Science, Prague 1967* (Academia Publication House of the Chechoslovakian Academy of Science, Prague), pp. 563–567.

Owren, M., Hopp, S., Sinnott, J., and Petersen, M. (**1988**). "Absolute auditory thresholds in three Old World monkey species (*Cercopithecus aethiops, C. Neglectus, macaca fuscata*) and humans (*homo spaiens*) ," J. Comp., Psych. **102**, 99–107.

Pisoni, D. (**1977**). "Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in

stops," J. Acoust. Soc. Am. **61**, 1352–1361.
Pisoni, D., Aslin, R., Perey, A., and Hennessy, B. (**1982**). "Some effects of training on identification and discrimination of voicing contrasts in stop consonants," J. Exp. Psychol. Hum. Percept. Perform. **8**, 297–314.
Pisoni, D., Lively, S., and Logan, J. (**1994**). "Perceptual learning of speech contrasts: Implications for theories of speech perception," in *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words*, edited by J. Goodman and H. Nusbaum, pp. 121–166 (MIT Press, Cambridge, MA).
Sinnott, J. (**1994**). "Comparisons of human and monkey differential sensitivity to speech, non-speech, and monkey sounds," in *Current Topics in Acoustical Research, Vol. 1*, edited by J. Menon, pp. 355–363 (Council for Scientific Research, Trivandrum, India).
Sinnott, J., and Adams, F. (**1987**). "Differences in human and monkey sensitivity to acoustic cues underlying voicing contrasts," J. Acoust. Soc. Am. **82**, 1539–1547.
Sinnott, J., and Saporita, T. (**2000**). "Differences in American-English, Spanish and monkey perception of the /say-stay/ trading relation," Percept. Psychophys. **62**, 1312–1319.
Streeter, L. (**1976**). "Language perception of 2-month old infants shows effects of both innate mechanisms and experience," Nature (London) **259**, 39–41.
Streeter, L., and Landauer, T. (**1976**). "Effects of learning English as a second language on the acquisition of a new phonetic contrast," J. Acoust. Soc. Am. **59**, 448–451.
Waters, R., and Wilson, W. (**1976**). "Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants," Percept. Psychophys. **19**, 285–289.
Williams, L. (**1977a**). "The perception of stop consonant voicing by Spanish-English bilinguals," Percept. Psychophys. **21**, 289–297.
Willams, L. (**1977b**). "The voicing contrast in Spanish," J. Phonetics, **5**, 169–184.

Release from speech-on-speech masking by adding a delayed masker at a different location

Brad Rakerd[a)]
Department of Audiology and Speech Sciences, Michigan State University, East Lansing, Michigan 48824

Neil L. Aaronson and William M. Hartmann
Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824

(Received 27 November 2004; revised 4 December 2005; accepted 5 December 2005)

The amount of masking exerted by one speech sound on another can be reduced by presenting the masker twice, from two different locations in the horizontal plane, with one of the presentations delayed to simulate an acoustical reflection. Three experiments were conducted on various aspects of this phenomenon. Experiment 1 varied the number of masking talkers from one to three and the signal-to-noise (S/N) ratio from −12 to +4 dB. Evidence of masking release was found for every combination of these variables tested. For the most difficult conditions (multiple maskers and negative S/N) the amount of release was approximately 10 dB. Experiment 2 varied the timing of leading and lagging masker presentations over a broad range, to include shorter delay times where room reflections of speech are rarely noticed by listeners and longer delays where reflections can become disruptive. Substantial masking release was found for all of the shorter delay times tested, and negligible release was found at the longer delays. Finally, Experiment 3 used speech-spectrum noise as a masker and searched for possible energetic masking release as a function of the lead-lag time delay. Release of up to 4 dB was found whenever delays were 2 ms or less. No energetic masking release was found at longer delays.
 [DOI: 10.1121/1.2161438]

PACS number(s): 43.66.Pn, 43.66.Qp, 43.55.Hy [AK] Pages: 1597–1605

I. INTRODUCTION

When several people talk at once, listening to just one of them can be difficult. The extent of the difficulty depends upon a number of factors, including the spatial arrangement of the talkers, the similarity of their voices, and the content of the messages (Cherry, 1953; Moray, 1959; Yost, 1997; Bronkhorst, 2000). It might be expected that the difficulty could only increase if the distracting talkers' speech power was to be doubled by an acoustical reflection, but Freyman and colleagues have recently found evidence to the contrary (Freyman *et al.* 1999, 2001, 2004).

In a baseline condition, Freyman *et al.* mixed together independent speech samples from multiple talkers and presented the combined signals from a single loudspeaker directly in front of a listener. The listener's task was to attend to the message of a single target talker and to ignore the competing messages of any other talkers, which were distractors. The surprising finding was that this task became easier when an added copy of the distractors' speech was presented from a second loudspeaker off to the listener's right side, with the copy shifted in time by a few milliseconds. This manipulation produced perceptible changes in the spatial characteristics of the distracting speech, and a substantial reduction in its interfering effects. An important further finding was that the reduction appeared to be based largely on release from informational masking (Watson *et al.*, 1976; Leek *et al.*, 1991; Kidd *et al.*, 1994), as evidenced by the fact that it was obtained with speech distractors but not with other physically comparable sounds.

Most experiments conducted to date have shifted the distractors 4 ms forward in time, so that their presentation leads at the right side relative to the front where the target is located (Freyman *et al.* refer to this as condition F-RF). A shift of this kind can be expected to elicit a robust precedence effect for speech, with its leading and lagging copies fused into a single perceptual image that is localized at or near the location of the leading sound source (Wallach *et al.*, 1949; Litovsky *et al.*, 1999). Specifically, the image of the distractors should be localized at or near the right-hand loudspeaker (+60°) and well away from the target speech location (0°). Freyman *et al.* report just such an effect.

Significant masking release has also been found for a condition where the distractor speech is shifted to lag behind by 4 ms at the side. Freyman *et al.* report that in this condition (their F-FR condition) the distractors should be localized slightly to the right of the front speaker, consistent with a model of the precedence effect (Shinn-Cunningham *et al.*, 1993), and that a small perceptual separation of this kind may provide a basis for differentiating the target and distractor messages. Also, binaural disparities associated with the two-source presentation of the distractors create a relatively diffuse auditory image in this condition. By contrast, the target speech which is presented from a single loudspeaker straight ahead has a more compact image, and this difference in the images may aid a listener in attending to the target.

[a)]Author to whom correspondence should be addressed; electronic mail: rakerd@msu.edu

The primary motivation for the present study was to extend the investigation of this masking release phenomenon to a much broader range of time shifts than have been examined to date. The precedence effect generally operates in rooms, where listeners experience great variation in the timing associated with acoustical reflections. There are differences among rooms owing to their differing physical dimensions, and differences within a room owing to variations in the positioning of talkers and listeners. Our goal in the present study was to represent this reflection delay time variation, and to examine its consequences for speech masking release. Accordingly, perceptual tests were done here at delays comparable to those that have been examined previously (±4 ms), at delays much briefer than this, and at delays much longer.

The speech stimuli used in this study were drawn from the Coordinate Response Measure (CRM) speech stimulus set (Bolia *et al.*, 2000). This corpus is well suited to the study of speech masking and masking release (Brungart, 2001; Brungart and Simpson, 2002; Brungart *et al.*, 2001, 2005). With the CRM, it is possible to create a competition among two or more messages, all spoken by different talkers, and to precisely manipulate the relative intensities of the various talkers' speech. These were the principle variables examined in Experiment 1. Based on the outcome of that experiment, a detailed examination of delay times and speech masking release was then conducted in Experiment 2. Finally, in Experiment 3, speech-spectrum noise was used to look for energetic masking release effects at multiple delay times.

II. EXPERIMENT 1: THE NUMBER OF DISTRACTING TALKERS AND S/N

A computer program and graphical user interface were written to assess speech understanding with the CRM as a function of any of several parameters that have been shown to be relevant to the study of speech masking release. These included the number of distracting talkers present, the delay associated with a multisource presentation of the distractors' speech, and the relative levels of the speech produced by the distractor talkers and by a target talker, which dictates the signal-to-noise ratio (S/N). Experiment 1 examined release from speech masking as a function of the number of distracting talkers and as a function of S/N. All other parameters for the experiment were fixed at values selected to match those employed in previous studies.

A. Subjects

The subjects of this study were two women (ages 20 and 63) and two men (ages 24 and 51). All of the subjects had hearing thresholds within 20 dB of audiometric zero at speech frequencies. All of them were practiced listeners. Two of the subjects were among the authors of this study. The other two had no knowledge of the purpose of the experiment.

B. Stimuli

Stimuli were a subset of the sentences comprising the CRM corpus. This corpus features eight different talkers—four men and four women—each saying 256 different versions of the following sentence:

"Ready ⟨call sign⟩, go to ⟨color⟩⟨number⟩ now . "

The sentence versions differ according to the values assigned to the variables ⟨call sign⟩, ⟨color⟩, and ⟨number⟩. Stimulus files are approximately equal in overall length and time aligned at the onset of the word "Ready."

For the present study we used 64 sentence versions, as recorded by each of the four female talkers (CRM talkers 04-07). The 64 sentences included all possible combinations of four call signs (Charlie, Ringo, Laker, Hopper), four colors (blue, red, white, green), and four numbers (1, 2, 3, 4).

C. Anechoic room and loudspeaker arrangement

All tests were conducted in a 3.0-m-wide × 4.3-m-long × 2.4-m-high anechoic room (IAC 107840). A subject chair was placed at approximately the middle of this room. Two loudspeakers were also placed in the room. One of the loudspeakers was directly in front of the subject chair and 1.5 m away. The other speaker, also 1.5 m away, was 60° to the right. The loudspeakers were set at the approximate ear height of a seated subject. Subjects always sat facing straight ahead.

Stimulus presentation. On test runs referred to here as *Front-Only*, the target speech and the distractors were mixed and presented exclusively from the front loudspeaker location. This was the location that subjects were asked to monitor throughout.

On runs referred to as *Front+Right* the target and distractors were presented from the front speaker, as for Front-Only runs, and in addition the distractors were time shifted and presented from the right-hand loudspeaker.[1] The time-shift associated with the right speaker presentation in this experiment was fixed at +4 ms (where + indicates right speaker leading), to agree with a number of previous studies and to invite precedence effect capture of the distractors and their localization off to the subject's right-hand side.

D. The task

Subjects were instructed to monitor the front loudspeaker throughout testing, always listening for sentences that featured the call sign *Laker*. A sentence with this call sign was presented at the front location on every trial. It could be spoken by any of the four female talkers, and it could feature any of the four colors and any of the four numbers. All of these varied randomly from trial to trial. The subject's task on each trial was to monitor for the target call sign and then to report the color/number combination in the sentence spoken by the talker who addressed *Laker*. Judgments were reported via a custom-designed response box.

Testing was done in runs of 35 trials each. The first 5 trials were treated as practice (without feedback). A percent correct score was calculated based on performance over the remaining 30 trials of the run, with a correct response requir-

ing both the correct color and the correct number. The chance rate for performance of this task is one in sixteen (6.25%).

E. Number of distractors and S/N

In all, there were 24 test conditions in the experiment, based on parameter settings for the number of distractors and for S/N as detailed in the following, and on whether the stimulus presentation was Front-Only or Front+Right. Settings on these parameters were fixed within a test run, and varied randomly across the runs. Each subject completed a total of three runs for each combination of the settings. The ordering of these runs was random and different for every subject.

Prior to the start of formal testing, subjects were given several sample runs to familiarize them with the testing procedures. They received no feedback about response accuracy on these sample runs.

1. Number of distractors

The number of distractor talkers competing with the target talker on a test run was set at one, two, or three. A new set of sentences was generated at the start of each trial of the run, based on this parameter. First, the four female talkers available in CRM were randomly assigned to the target sentence and to the distractor sentences, with the constraint that each talker could be used only once. Next, colors and numbers were randomly assigned to the different sentences, also with the constraint that they could be used only once. Finally, the appropriate sentences were accessed from the CRM database and level adjusted, time shifted, and digitally mixed, as called for. The resulting signals were delivered via a two-channel D/A converter (sample rate=40 ksps), amplifier, and single-driver loudspeakers (Minimus 3.5).

2. S/N ratio

The rms presentation level of individual distractor sentences was fixed in the experiment at 65 dB SPL. Whenever one, two, or three distractor sentences were presented on a test, each sentence was presented at this level. Whenever leading and lagging copies of a distractor sentence were presented from the front and right-hand loudspeakers, both copies were presented at this level. S/N ratio was manipulated by varying the level of the target sentence presentations relative to the fixed level of the individual distractors. Across conditions, targets were presented at levels of 53, 57, 61, 65, and 69 dB, yielding S/N ratios of −12, −8, −4, 0, and +4 dB.[2] For the one-distractor test conditions, runs were conducted at the four lowest values of S/N (−12, −8, −4, and 0 dB). For the two- and three-distractor tests, runs were done at the four highest values (−8, −4, 0, and +4 dB).

F. Results and discussion

Figure 1 shows the results of Experiment 1. Percent correct scores (based on a total of 90 trials per condition per subject completed over three runs) are reported as a function of S/N ratio. There are separate plots for the Front-Only (open symbols, dashed line) and Front+Right (closed symbols, solid line) conditions. Data points are for the average subject ($n=4$). Error bars correspond to the standard deviation over subjects. Results are broken out in different panels according to the number of distractors.

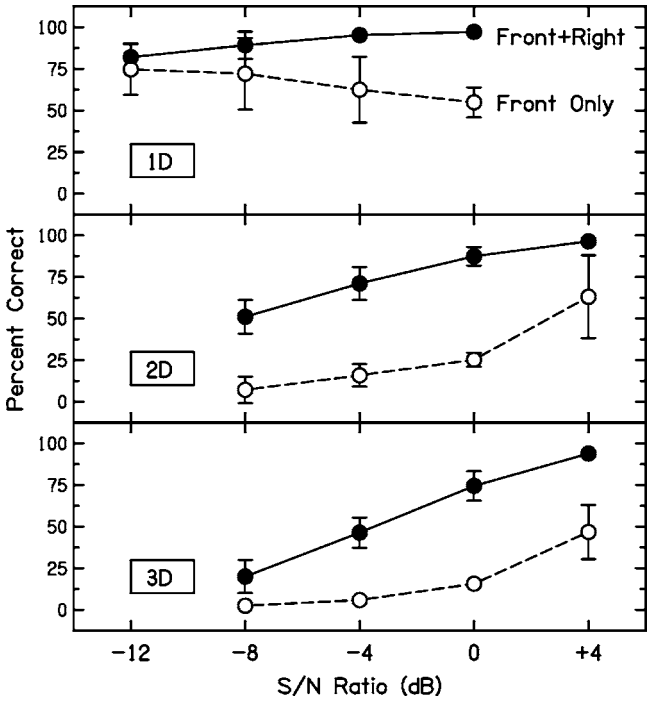


FIG. 1. Results of Experiment 1. Percentage of correct responses to a target presented from a front speaker for two conditions: *Front Only*, where distractors were presented from the same front speaker as the target; and *Front+Right* where distractors were presented from the front speaker and again from a right-hand speaker, with the right leading by 4 ms. The mean score and standard deviation over subjects are plotted as a function of S/N ratio. Top, middle, and bottom panels show results for interference from one distracting talker (1D), two distractors (2D), and three distractors (3D), respectively. Error bars are 2 s.d. in overall length.

1. One distractor

The top panel shows the results for the case of one distractor. There are two findings of note. The first is that Front+Right presentation was found to afford an advantage over Front-Only presentation at every S/N ratio. Hence, with one distractor there was consistent evidence of release from masking for speech.

Also notable is the finding that with one distractor the Front-Only function had a negative slope; that is, performance was inversely related to S/N. For example, at a S/N of 0 dB subjects correctly identified the target color/number combination on 55% of trials, but when the target speech level was reduced by 12 dB (S/N=−12), performance improved significantly to 75% correct [$t(3)=5.15; p<0.02$]. Brungart (2001) also found instances of reversal of this kind, as well as instances of plateauing where performance held constant as S/N was reduced over a considerable interval. Brungart (2001) reports that both such effects can occur on tests where: (a) a listener must attend to one talker's speech and ignore that of a competing talker with similar vocal characteristics; and (b) S/N is reduced over the interval from 0 dB to approximately −10 dB (Egan *et al.*, 1954; Dirks and Bower, 1969; Freyman *et al.*, 1999). The level difference

between the two talkers' speech is a cue that aids a listener in differentiating the messages within this regime. The size of that level difference increases as S/N decreases below 0 dB. So long as the S/N degradations do not become too severe, this strengthening level-difference cue can support good listener performance.

2. Two distractors

Results for two distractors are shown in the middle panel of Fig. 1. They show a very clear advantage for two-source presentation at all S/N ratios tested (−8 to +4 dB), with performance for Front+Right conditions exceeding comparable Front-Only conditions by as much as 62 percentage points (two distractors, S/N=0 dB).

The results for two distractors do not show the negative S/N effect for Front-Only presentation that was found with one distractor. Brungart *et al.* (2001) reported a similar result. Reversal/plateauing effects clearly present in their data for one distracting talker were not found with two or three distractors. Apparently when the number of distractors exceeds one a listener can no longer reliably distinguish the target talker's message based on overall level difference. Instead, the limiting factor becomes S/N.

The S/N values corresponding to 50%-correct performance of this task were approximately +3.0 dB for Front-Only presentation and −8.0 dB for Front+Right presentation. Hence, repeated presentation of the distractors afforded 11 dB of masking release in the two-distractor condition. This compares with an estimated release of 10 dB reported by Freyman *et al.* (2001) for tests done with two distractors.

3. Three distractors

Results for three distractors are shown in the bottom panel. They are similar to those for two distractors, with performance reduced somewhat overall. For the Front-Only condition, performance fell to near the chance-level floor (6.25%) for S/N's less than 0 dB. A comparison of the 50% points of the functions (extrapolated for F-Only) shows a masking release of approximately 8 dB with three distractors (+4.5 for F-Only vs −3.5 dB for Front+Right), which is 3 dB less than the release found here with two distractors. This difference between the two and three distractor results is consistent with a recent report by Freyman *et al.* (2004), who found maximal masking release with two distractors and progressively decreasing release thereafter as the number of distractors was increased from three to ten.

4. Overall

Overall, the results of Experiment 1 showed robust evidence of release from speech masking whenever masker signals were presented simultaneously from two locations in the horizontal plane—one front and one right—with the right-hand presentation shifted ahead in time by 4 ms to trigger a strong precedence effect.

III. EXPERIMENT 2: DIFFERENT DELAYS

Previous investigations of release from speech masking, including Experiment 1 of this paper, have employed a distractor delay time constant of 4 ms, which can be expected to elicit a substantial precedence effect for speech. Experiment 2 examined a broad range of delays, to include this 4 ms value as well as a number of values substantially briefer than this and a number substantially longer. Following Freyman *et al.* (1999), the experiment tested both positive delays, where the distractor speech was time shifted to lead in the right-hand speaker relative to the front, and negative delays, where the distractors were shifted to lag in the right-hand speaker.

A. Methods

1. Subjects

There were six subjects in this experiment. Four of the subjects (S1–S4) were the same individuals who participated in Experiment 1. They completed all testing for that experiment before beginning this one. Two new subjects (S5 and S6), one male (age 22), one female (age 24), were tested here as well. Both of the new subjects were experienced listeners. Both had hearing thresholds within normal limits. Neither had any knowledge of the purpose of this study.

2. Stimuli

As in Experiment 1, the target (call sign Laker) was always presented from the front speaker together with distractors, and an additional copy of the distractors was optionally presented from the right-hand speaker with an intervening time delay. The number of distractors and S/N were both fixed for this experiment, based on the results of Experiment 1. The number of distractors was fixed at two and the S/N was fixed at −4 dB. In Experiment 1, this combination was found to yield a substantial separation in performance between the Front-Only (16% correct) and Front+Right (71% correct) conditions, with no evidence of either floor or ceiling effects.

3. Delays and test runs

Front+Right test runs were done in Experiment 2 at each of 16 different delays. The delays tested were −64, −32, −16, −4, −0.5, −0.25, 0.0, +0.25, +0.5, +1, +2, +4, +8, +16, +32, and +64 ms, where negative delays correspond to time shifts in which the right-hand loudspeaker presentation lagged relative to the front speaker, and positive delays to shifts where the right-hand loudspeaker led. In all, subjects completed a total of three runs at each delay. They also completed a set of three runs in a Front-Only condition to establish a baseline comparison for this experiment. The order of these runs was random, and different for every subject.

B. Results and discussion

1. Front-Only tests

Figure 2 shows the results of Experiment 2. In Front-Only testing with two distractors and a −4-dB S/N, the six subjects averaged 15.9% correct. The hatched area in Fig. 2 brackets a 95% confidence interval around this mean score for comparison with the various Front+Right conditions.

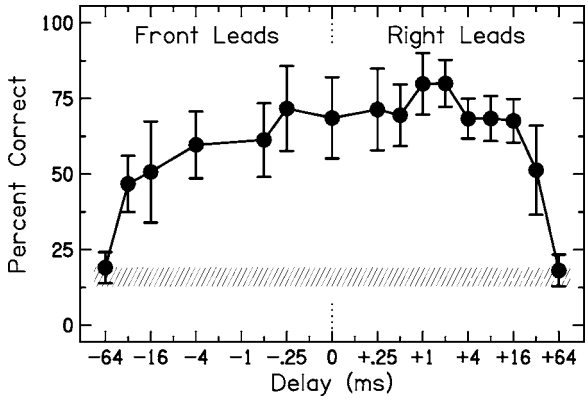


FIG. 2. Results of Experiment 2. Percent-correct mean and standard deviation over subjects, plotted as a function of the delay between the front and right loudspeakers. For negative delays, the distractors' speech led at the front speaker, i.e., at the target speaker. For positive delays it led at the right speaker. The hatched area shows a 95% confidence interval around the mean score (15.9%) when stimuli were presented from the front speaker only.

2. Front + Right tests

Results for the Front+Right tests are plotted as a function of delay. Data points correspond to the subjects' mean percent correct score. Error bars show ±1 s.d. over subjects. Performance with Front+Right presentations fell into two regimes, depending on the duration of the delay.

For the two longest delays tested, one forward in time (+64 ms) the other backward (−64 ms), there was no evidence of release from masking, with scores in both instances falling well within the 95% confidence interval for Front-Only presentations.

For all of the remaining tests, which were done at shorter delays ranging from −32 to +32 ms, there was consistent evidence of release from masking for speech. Scores were substantially greater than for the Front-Only condition in every instance, always falling well outside the Front-Only 95% confidence interval.

An analysis of variance on the results of Experiment 2 showed a significant difference among the conditions overall $[F(16,80)=58.86; p<0.001]$. Paired comparisons (Bonferroni-protected, $p<0.05$ experimentwise) between the Front-Only condition and each of the Front+Right conditions with time delays between −32 and +32 ms found a significant difference in every case.

C. Conclusions

The present findings support several new conclusions regarding speech masking release elicited by adding a delayed masker at a different location in the horizontal plane.

1. A wide-ranging effect

The first conclusion is that the effect is quite general. Significant release from masking was seen here for delays brief enough to fall within the regime of summing localization, where the leading copy of the distractors and the lagging copy both contribute substantially to perceived location. Summing localization occurs for delays less than 1 ms (Blauert, 1971). The present study found masking release for five such delays (−0.5, −0.25, 0, +0.25, and +0.5 ms). Significant masking release was also found here for a family of longer delays for which the precedence effect is fully engaged and sound localization depends mostly on cues associated with the leading presentation only (Wallach *et al.*, 1949; Litovsky *et al.*, 1999). Altogether, the set of delays found to exhibit significant release from speech masking in this study ranged from −32 to +32 ms. This spans the full range of acoustical reflection delay times that a listener is likely to encounter on an everyday basis in rooms.

2. Release at negative delays

A second conclusion to be drawn from the results of Experiment 2 is that masking release for speech is substantial for negative delays where the repeated distractors lag the distractors presented with the target. The effect for negative delays was found to reach out to at least −32 ms. This confirms and extends the report of Freyman *et al.* (1999), who found significant masking release at −4 ms. They pointed out that at −4 ms listeners may have been aided in distinguishing targets from distractors by a relative difference in diffuseness of the images, with images of the distractors being the more diffuse. In the present study, diffuse distractor images were also noted by listeners at negative delay times.

A temporal asymmetry. Figure 2 shows some evidence of an asymmetry between percent-correct scores obtained with negative delays and those obtained with positive delays, with the former generally somewhat lower. For example, the subjects' mean percent-correct score was 8.6% lower at −4 ms than at +4 ms (59.7% vs 68.3%), and it was 16.9% lower at −16 ms than at +16 ms (50.7% vs 67.6%).[3] We examined the individual subject records to learn more about the details of this asymmetry.

Figure 3 shows each subject's percent-correct scores for the Front+Right tests, plotted as a function of delay. Also shown, for comparison, is the score obtained in the baseline Front-Only condition (the hatched area). The error bars (and the 95% confidence interval for Front-Only) are based on a subject's standard deviation over test runs (3 runs per condition). Quintic functions, indicated in the figure by the solid-line plots, were fitted to each subject's delay series. These functions accounted for over 90% of the variance in the individual subject data. The area under each function was integrated over all negative delay times and over all positive delays to compare the strengths of the perceptual effects in these two regimes.

For one subject (S1), the negative and positive areas differed by less than 1%. But for the other subjects (S2 through S6) the areas differed by 5% or more, and in every case the negative area was smaller than the positive area. A statistical comparison based on the results for all six subjects showed the negative and positive areas under the functions to be significantly different $[t(5)=3.77;\ p<0.01]$, with the negative area 8.8% smaller on average. It appears that negative delay times, although effective at eliciting significant masking release relative to baseline (Front-Only), are nevertheless somewhat less effective than the corresponding positive delays.

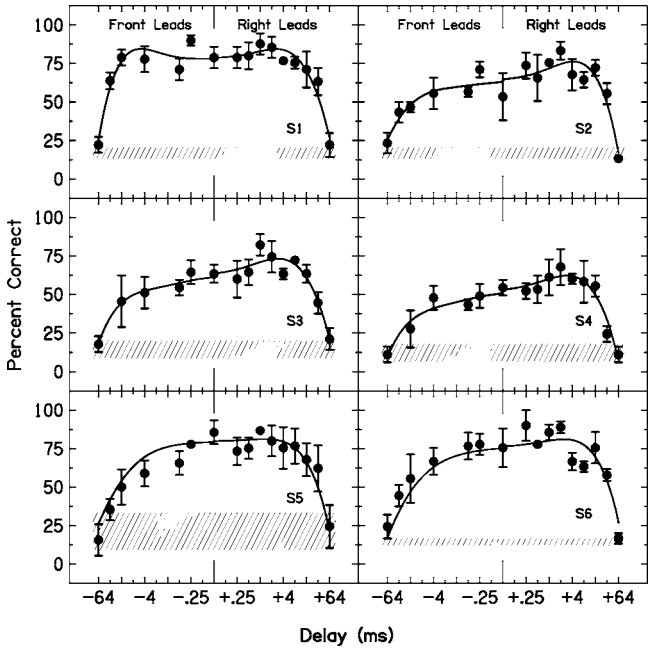


FIG. 3. Individual subject results for Experiment 2. Percent-correct mean and standard deviation over runs ($n=3$), plotted as a function of the delay between the front and right loudspeakers, and fitted with a best-fit fifth-order function (solid line). The hatched area shows a 95% confidence interval around the mean score when stimuli were presented from the front speaker only.

3. No release at very long delays

Sixteen different delay times were tested in this experiment and evidence of significant masking release was found for all but two of them. For both of the failing conditions (−64 and +64 ms) there was a 64-ms time delay between leading and lagging presentations of the distractors' speech. If acoustical reflections of speech are delayed by 64 ms in a room there is a good probability that listeners will become aware of them as echoes, so much so that the reflections may become disruptive (Haas, 1951; Kutruff, 1979, pp. 174–177). The present results suggest that once speech reflections become noticeable in this way they can no longer participate in speech masking release.

IV. EXPERIMENT 3: SPEECH-SPECTRUM NOISE MASKERS

In Experiments 1 and 2, masking release was found for a number of tests in which speech maskers were presented from two locations in the horizontal plane (front and right) with an intervening time delay. This masking release could have taken either or both of two forms. First, it is possible that there was release from energetic masking (EM), which takes place at levels as low as the auditory periphery. Second, alternatively or additionally, there could have been release from informational masking (IM), which takes place more centrally.

Freyman *et al.* (1999) conducted an extensive search for EM release. They fixed masker-signal delay at each of two values, +4 and −4 ms, and then replaced their speech maskers with continuous noise. This noise had a power spectrum that resembled the speech, and was therefore subject to a similar level of EM release if any was present. The experiment found no evidence of EM release, however. To the contrary, subjects performed somewhat worse with the noise in the Front+Right condition than in the Front-Only condition. Virtually all of the release observed with speech maskers in −4 and +4-ms tests could therefore be attributed to informational masking release only.

Experiment 2 of the present study found significant release with speech maskers for the delay times of −4 and +4 ms and for a number of additional delays that had not been previously tested. Experiment 3 was therefore conducted to see whether EM release could explain any part of the results obtained at these new delay times.

A. Methods

1. Matched noise maskers

The maskers for this experiment were 192 noise files, created to individually match each of the 192 CRM speech files that served as maskers in Experiments 1 and 2. For each CRM file, we created a noise file equivalent by: (a) Fourier transforming the entire original signal as a single function; (b) randomizing the phases of all components; and (c) inverse Fourier transforming the derived spectrum to generate the noise waveform. The resulting noise signals matched the CRM originals in duration, power, and long-term amplitude spectrum. Lacking were characteristic speech temporal envelopes and any speech-like information. Everyone agreed that the maskers sounded like a swarm of bees.

2. Subjects

There were four subjects in the experiment. All four (S1, S2, S4, S6) had previously participated in Experiment 2, and all but one (S6) had participated in Experiment 1 as well.

3. Procedure

The noise masker test was modeled on the two-distractor speech masker test described earlier. The methods were, in fact, identical except that the two randomly selected CRM speech masker files were replaced by their corresponding noise masker files on each trial. S/N for this experiment was set at −10 dB, based on pilot testing which showed that at this level subjects performed the baseline Front-Only test better than chance and also at a low enough level to allow headroom for a search for possible energetic masking release.

4. Front-Only runs

A set of baseline runs was conducted in the Front-Only condition with the number of distractors set at two, and the S/N set at −10 dB. Also, for a second point of reference, Front-Only runs were done with the S/N increased by 4 dB compared to baseline (S/N=−6 dB).

5. Front + Right runs

For the Front+Right condition, test runs were done at nine different values of time delay, covering the range from −32 to +32 ms where significant speech-masker release was

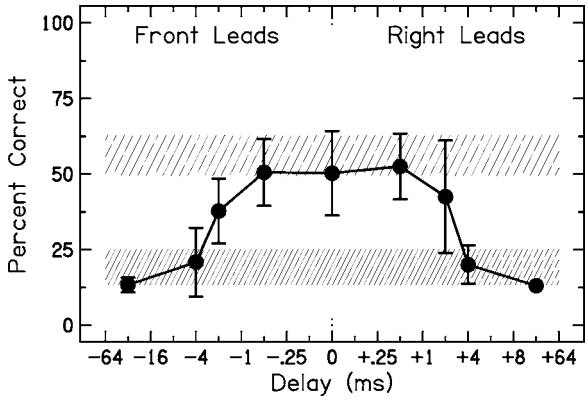


FIG. 4. Results of Experiment 3. The masker signals were noises with long-term spectra matched to the speech of Experiments 1 and 2. Closed circles connected by a solid line give percent-correct mean and standard deviation over subjects, as a function of the delay between the front and right loudspeakers. The hatched areas show 95% confidence intervals around the mean scores obtained when stimuli were presented from the front speaker only at a baseline level (S/N=−10 dB; hatched area at the bottom) and with the target signal level increased by four dB (S/N=−6 dB; hatched area at the top).

found in Experiment 2. Specifically, tests were done at delays of −32, −4, −2, −0.5, 0.0, +0.5, +2, +4, and +32 ms. Subjects completed three runs (90 total trials) for each test condition, with run order randomized differently for each subject.

B. Results and discussion

The results of Experiment 3 are given in Fig. 4. The lower hatched area shows a 95% confidence interval based on the four subjects' performance in the baseline Front-Only condition (−10-dB S/N). The upper hatched area gives a confidence interval around the mean score for the second Front-Only test, in which S/N was increased by 4 dB (S/N =−6 dB). Results for the Front+Right condition tests are plotted as a function of delay, with closed circles connected by a solid line.

Freyman *et al.* found no evidence of EM release for the delay times of −4 and +4 ms. The present results confirm that finding. We also found no EM release at the two longer delay times tested here, −32 and +32 ms. In all of these instances, the subjects' mean percent correct score fell well within or below the confidence interval for baseline Front-Only presentations (−10-dB S/N). It follows that when delay times are several milliseconds long or longer, in either a positive or a negative direction, subjects experience informational masking release only.

A very different pattern of results emerged here for short delays. For all five of the short delays tested (−2, −0.5, 0.0, +0.5, and +2 ms), the mean percent-correct score was above the Front-Only baseline, and in all but one case (+2 ms) the error bars were above it as well. An analysis of variance on these results was significant overall [$F(5,60)=12.91$; $p<0.001$], reflecting significant pairwise differences between the Front-Only baseline score and the scores obtained at −2, −0.5, 0.0, and +0.5 ms (Bonferoni-protected paired comparisons, $p<0.05$). A separate statistical comparison with the Front-Only condition in which S/N had been increased by 4 dB (the −6-dB S/N test), found the Front+Right mean score for the delay time of −2 ms to be significantly lower and the remaining scores to be not significantly different ($p>0.05$).

Altogether, these results indicate: (a) that there is consistent EM release for speech maskers when time delays are brief (2 ms or less); and (b) that the resulting masking release can be equivalent to an increase in signal level of 4 dB.

1. Individual differences

The EM release effects described here were, for the most part, very consistent across listeners. The only notable individual differences were at the delay times of −2 and +2 ms. Three of the four listeners showed diminished but still measurable release at one or both of these delays relative to the release seen at briefer delays (−0.5, 0.0, and +0.5 ms). For the fourth subject (S6) there was no diminution. Release at ±2 ms was the same as at the other brief delays.

2. Cause of release from energetic masking

Conjectures as to the origin of the EM release seen here begin by recognizing that when the masker is presented simultaneously from two different sources with one source delayed, and when the sources are differently located and the signals are diffracted around the head, there is an effective filtering of the masker. Therefore, we performed a calculation of the overall masker transfer functions at the two ears. Calculations were done using a spherical head model as described by Kuhn (1977) and implemented by Constan and Hartmann (2003). The model describes diffraction of plane waves around the head and automatically includes frequency-dependent interaural timing and level differences. For each ear, computations of complex transfer functions were made independently for front and right-hand sources. An additional delay ($\tau=+0.5$, 0.0, or −0.5 ms) was added to the right source, and the complex functions for the two sources were summed to obtain the transfer function at the ear.

Figure 5 shows the amplitude response functions (magnitude of the complex functions) for the left and right ears, depending on the value of τ (top panel: $\tau=+0.5$ ms; middle: $\tau=0.0$ ms; bottom: $\tau=-0.5$ ms). The low-frequency limit is 2.0 as expected, corresponding to the addition of two identical signals at a point in free field located at the center of the head. The low-frequency functional form is consistent with delay-and-add filtering where the delay is the sum of τ and a head-related delay. These amplitude response functions lead to the first conjecture about the origin of EM release. The functions show deep spectral valleys in the vicinity of speech formants caused by the interference effects from the speaker on the right. It is possible that the appearance of these valleys in the masker enabled listeners to hear out important features of the target talker. Further, the valleys at the left ear were normally different from those at the right, giving listeners twice the opportunity to hear out target features.

Figure 6 shows the interaural phase response, which leads to a second conjecture. It is possible that release from EM resulted from a binaural masking level difference of the

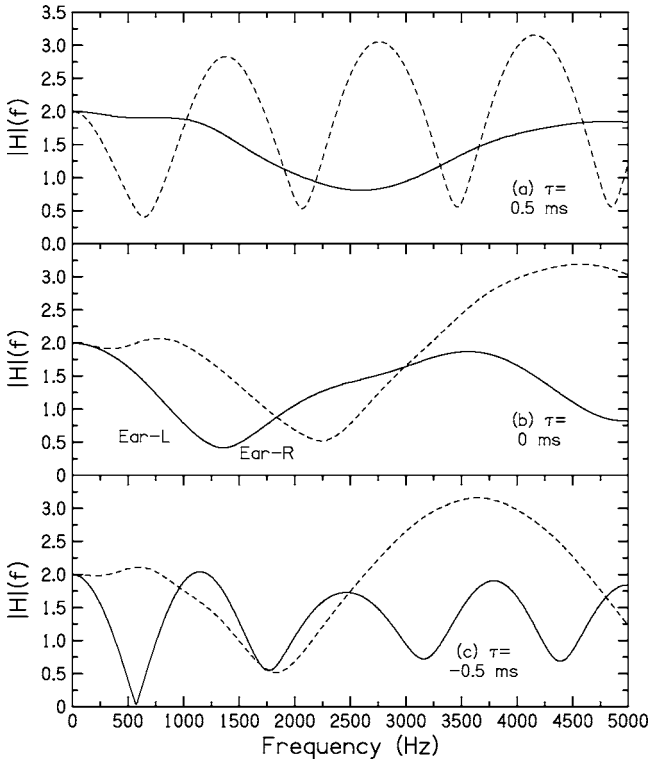


FIG. 5. Filter amplitude transfer functions for the masker delivered from two separated sources with time delay of +0.5, 0.0, or −0.5 ms, corresponding to test conditions where release from energetic masking was observed.

form $N_\phi S_0$. The figure shows that the interaural phase (Φ) of the maskers is near ±180° in a number of frequency regions, including the important region near 500 Hz, where binaural sensitivity is greatest. An interaural phase of ±180° corresponds to the condition $N_\pi S_0$.

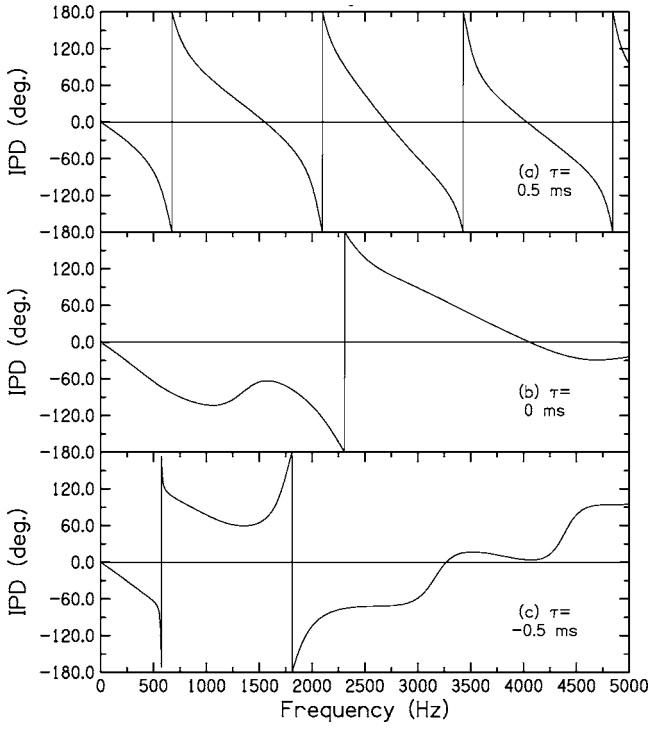


FIG. 6. Filter interaural phase response for the masker delivered from two separated sources with time delay of +0.5, 0.0, or −0.5 ms, corresponding to test conditions where release from energetic masking was observed.

Similar calculations for values of delay greater than ±0.5 ms show similar structure in both the amplitude responses and the interaural phase responses but the structure varies more rapidly with frequency. The above-noted conjectures would continue by suggesting that as the delays increase to 2.0 ms and beyond, the spectral or interaural features of the masker are confined to narrow, disconnected frequency regions that offer minimal advantage. Further investigation of these conjectures is beyond the scope of this report.

3. EM release and IM release

EM release of up to 4 dB was found at several different short delay times in this experiment. These effects, though consistent, were relatively small compared to the release seen in Experiment 1 with comparable speech maskers. There, we found 11 dB of masking release on a test done with two competing talkers, and 8 dB of release on a test with three talkers. It seems likely, therefore, that in speech-on-speech masking situations listeners will experience two kinds of masking release at very short delay times. There will be an EM release component of up to 4 dB, and an IM component that may be as large or larger.

V. SUMMARY

The CRM speech corpus was used here to test for release from speech masking when the masker was presented from two locations in the horizontal plane (front and right), with an intervening time delay. Key variables were (a) the number of distractor talkers who competed with a target talker, (b) the S/N for target and distractors, and (c) the delay associated with a multi-source presentation of the distractors' speech. Two experiments employed CRM speech maskers. One fixed the delay time at +4 ms and varied the number of distractors (1, 2, or 3) and S/N (−12, −8, −4, 0, or +4 dB). The other fixed the number of distractors (2) and S/N (−4 dB) and varied the delay over a wide range (16 different values between −64 and +64 ms). A third experiment was conducted with noise maskers matched to the spectra of CRM speech stimuli. It searched for energetic masking release at a range of delay times. The experiments support the following conclusions.

(1) Masking release occurs in a wide variety of listening conditions. For the most challenging conditions, such as multiple distracting talkers with negative S/N, the amount of release is of the order of 10 dB.
(2) Masking release is robust over variations in the masker delay time. Significant release was found here for delays spanning a wide range (−32 to +32 ms), consistent with the wide variation in acoustical reflection delays that is present in everyday rooms.
(3) Significant masking release takes places at negative delays, where the repeated distractors lag the distractors presented with the target. The diffuseness of distractor images may provide a cue that aids listeners in these instances.
(4) The upper limit on delays that elicit masking release falls somewhere between 32 and 64 ms. In 1927, Petzold de-

fined the "threshold of masking" as a delay for which a noticeable deterioration of the acoustical impression of speech occurs. The value was 50±10 ms. The conclusions of the celebrated paper by Haas (1951) confirmed the perceptual significance of 50 ms. As noted by Blauert (1983, p. 226), "At delay times less than 50 ms, echoes are no longer perceived as annoying even if the reflection is considerably stronger than the primary sound." The results of Experiment 2, wherein masking release was seen for a delay of 32 ms but not seen for a delay of 64 ms, are consistent with the idea that release from speech-on-speech masking fails when the delay of the delayed masker is long enough that the delayed masker no longer fuses with the primary masker and becomes annoying.

(5) For delay times longer than 2 ms, virtually all of the masking release that takes place for speech can be attributed to release from informational masking. For delay times of 2 ms or less, there is also a component of release from energetic masking.

ACKNOWLEDGMENTS

We thank Associate Editor Armin Kohlrausch and three anonymous reviewers of this manuscript for their many useful comments. We also thank the authors of the CRM for sharing this very useful stimulus set. This research project was supported by Grant No. DC00181 from the NIDCD of the NIH.

[1]The Front-Only presentation condition of the present study corresponds to a condition that Freyman *et al.* (1999) refer to as Front-Front (F-F). Our Front+Right condition corresponds to their Front-Right/Front condition (F-RF).

[2]The noise level reference for calculations of S/N (65 dB) conforms to a single distractor sentence presented from a single loudspeaker. Whenever multiple distractors were presented and/or whenever distractors were presented from two loudspeakers (front and right), noise power increased and S/N was effectively worsened.

[3]Both of these group differences were statistically significant [±4 ms: $t(5)=2.87$, $p<0.035$; ±16 ms: $t(5)=3.24$, $p<0.023$].

Blauert, J. (**1971**). "Localization and the law of the first wavefront," J. Acoust. Soc. Am. **50**, 466–470.

Blauert, J. (**1983**). *Spatial Hearing* (MIT, Cambridge, MA).

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acust. Acta Acust. **86**, 117–128.

Brungart, D. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Brungart, D., and Simpson, B. (**2002**). "Within-ear and across-ear interference in a cocktail-party listening task," J. Acoust. Soc. Am. **112**, 2985–2995.

Brungart, D., Simpson, B., Darwin, C., Arbogast, T., and Kidd, G. (**2005**). "Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task," J. Acoust. Soc. Am. **117**, 292–304.

Brungart, D., Simpson, B., Ericson, M., and Scott, K. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech with one and two ears," J. Acoust. Soc. Am. **25**, 975–979.

Constan, Z. A., and Hartmann, W. M. (**2003**). "On the detection of dispersion in head-related transfer functions," J. Acoust. Soc. Am. **114**, 998–1008.

Dirks, D., and Bower, D. (**1969**). "Masking effects of speech competing messages," J. Speech Hear. Res. **12**, 229–245.

Egan, J. P., Carterette, E. C., and Thwing, E. J. (**1954**). "Some factors affecting multi-channel listening," J. Acoust. Soc. Am. **26**, 774–782.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.

Haas, H. (**1951**). "Uber den Einfluss eines Einfachechos auf die Horsamkeit von Sprache [On the influence of a single echo on the intelligibility of speech]," Acustica **1**, 49–58.

Kidd, Jr., G., Mason, C. R., Deliwala, P. S., Woods, W. S., and Colburn, H. S. (**1994**). "Reducing informational masking by sound segregation," J. Acoust. Soc. Am. **95**, 3475–3480.

Kuhn, G. F. (**1977**). "Model for interaural time differences in the azimuthal plane," J. Acoust. Soc. Am. **62**, 157–167.

Kuttruff, H. (**1973**). *Room Acoustics* (Applied Science, London).

Leek, M. R., Brown, M. E., and Dorman, M. F. (**1991**). "Informational masking and auditory attention," Percept. Psychophys. **50**, 205–214.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (**1999**). "The precedence effect," J. Acoust. Soc. Am. **106**, 1633–1654.

Moray, N. (**1959**). "Attention in dichotic listening: Affective cues and the influence of instructions," Q. J. Exp. Psychol. **11**, 56–60.

Petzold, E. (**1927**). *Elementare Raumakustik* (Bauwelt, Berlin), p. 8.

Shinn-Cunningham, B. G., Zurek, P. M., and Durlach, N. I. (**1993**). "Adjustment and discrimination measurements of the precedence effect," J. Acoust. Soc. Am. **93**, 2923–2932.

Wallach, H., Newman, E. B., and Rosenzweig, M. R. (**1949**). "The precedence effect in sound localization," Am. J. Psychol. **57**, 315–336.

Watson, C. S., Kelly, W. J., and Wroton, H. W. (**1976**). "Factors in the discrimination of tonal patterns. II. Selective attention and learning under various levels of stimulus uncertainty," J. Acoust. Soc. Am. **60**, 1176–1185.

Yost, W. A. (**1997**). "The cocktail party problem: Forty years later," in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson (Erlbaum, Hillsdale, NJ), pp. 329–347.

Effects of reverberation and masking on speech intelligibility in cochlear implant simulations

Sarah F. Poissant,[a] Nathaniel A. Whitmal III, and Richard L. Freyman
Communication Disorders Department University of Massachusetts Amherst, 6 Arnold House, Amherst, Massachusetts 01003

(Received 27 July 2005; revised 21 December 2005; accepted 5 January 2006)

Two experiments investigated the impact of reverberation and masking on speech understanding using cochlear implant (CI) simulations. Experiment 1 tested sentence recognition in quiet. Stimuli were processed with reverberation simulation (T=0.425, 0.266, 0.152, and 0.0 s) and then either processed with vocoding (6, 12, or 24 channels) or were subjected to no further processing. Reverberation alone had only a small impact on perception when as few as 12 channels of information were available. However, when the processing was limited to 6 channels, perception was extremely vulnerable to the effects of reverberation. In experiment 2, subjects listened to reverberated sentences, through 6- and 12-channel processors, in the presence of either speech-spectrum noise (SSN) or two-talker babble (TTB) at various target-to-masker ratios. The combined impact of reverberation and masking was profound, although there was no interaction between the two effects. This differs from results obtained in subjects listening to unprocessed speech where interactions between reverberation and masking have been shown to exist. A speech transmission index (STI) analysis indicated a reasonably good prediction of speech recognition performance. Unlike previous investigations, the SSN and TTB maskers produced equivalent results, raising questions about the role of informational masking in CI processed speech. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2168428]

PACS number(s): 43.66.Ts, 43.71.Ky, 43.66.Mk, 43.66.Dc [AJO] Pages: 1606–1615

I. INTRODUCTION

Since the introduction of the 3M House single-channel cochlear implant in 1980, cochlear implant technology has progressed rapidly. Several generations of internal devices, external components, and speech coding strategies have resulted in improvement of sentence recognition scores obtained in quiet from 0% with single-channel devices, to averages ranging from 70% to over 90% with current systems (Skinner *et al.*, 2002; Firszt *et al.*, 2004; Koch *et al.*, 2004; Spahr and Dorman, 2004). Although this excellent performance in quiet is extremely encouraging, it is important to also consider the ability of cochlear implant users to understand speech in real-world environments in which there is noise and reverberation. Performance has been shown to decrease significantly and to much greater degrees than in listeners with normal hearing when speech is presented in noisy backgrounds (Dorman *et al.*, 1998; Zeng and Galvin, 1999; Muller *et al.*, 2002; Firszt *et al.*, 2004; Koch *et al.*, 2004; Spahr and Dorman, 2004). Zeng and Galvin (1999) found that cochlear implant recipients required an 8–11-dB better S/N than listeners with normal hearing for equivalent (50% correct) performance on a task of vowel and consonant identification.

Much less known is the effect of reverberation, which is present in varying amounts in virtually all enclosed spaces, and which is likely to be a significant contributing factor to difficulties cochlear implant users experience in their daily lives. While a small amount of reverberation (e.g., reverberation times of 0.02 to 0.03 s) may benefit listeners with normal hearing (Lochner and Burger, 1961, 1964; Nabalek and Robinette, 1978), moderate and greater levels have a negative impact on speech understanding (Plomp, 1976). Such detrimental effects are exacerbated in the presence of hearing loss (Nabalek and Robinette, 1978; Irwin and McAuley, 1987; Helfer and Wilber, 1990). One reason for this may be related to the influence of temporal properties of the damaged auditory system. Dreschler and Leeuw (1990) found that gap detection ability (a measure of temporal resolution) for wideband noise was related to speech recognition in the presence of reverberation for listeners with hearing loss. Although it has been demonstrated that cochlear implant recipients generally have relatively normal temporal resolution abilities (Shannon, 1989, 1993), there is clearly reason to be concerned about the effects of reverberation on speech understanding in this population. This is because the information delivered through the cochlear implant is carried primarily by the temporal envelope in a small number of spectral channels. The dependence on temporal envelope information is so great that distortions of the envelope created by reverberation are likely to have a substantial impact on the speech information delivered by cochlear implants. The current investigation employs reverberation simulation using the image method (Allen and Berkley, 1979) which allows for strict control over the characteristics of the reverberant environments under study. This methodology provides several advantages over testing in naturalistic settings including (1) there is easy and precise manipulation of the acoustic characteristics of the room, (2) it allows us to study the impact of

[a]Electronic mail: spoissant@comdis.umass.edu

 0001-4966/2006/119(3)/1606/10/$22.50

reverberation without confounding noise and/or level effects, and (3) it allows for monaural as well as binaural testing in listeners with normal hearing.

This investigation uses simulations of cochlear implant listening (also referred to as vocoding or vocoders) to study the potential impact of reverberation, masking, and the combination of the two on the recognition of sentences. The simulation involves extracting the envelopes from several frequency bands and using the envelopes to modulate sine waves. Such simulations, as well as those using narrow-band noise carriers, have been employed for nearly a decade in experiments examining a variety of speech processor parameters (Shannon *et al.*, 1995; Dorman and Loizou, 1998; Fu *et al.*, 1998; Loizou *et al.*, 1999, 2000; Faulkner *et al.*, 2003; Qin and Oxenham, 2003). While the accuracy of any given cochlear implant simulation method is unknown, previous investigations have produced comparable results between such simulations and actual cochlear implant listening (Fu *et al.*, 1998; Friesen *et al.*, 2001; Fu and Galvin, 2001). Benefits of this type of modeling include the sheer number of subjects that can be tested, the control over precise processing characteristics, and the wide range of manipulable parameters. It has been repeatedly found that increasing the number of available spectral channels results in improved perception (Shannon *et al.*, 1995; Dorman *et al.*, 1997; Friesen *et al.*, 2001; Shannon *et al.*, 2004). Performance reaches asymptotic levels at different numbers of channels, depending upon the difficulty of the test material and/or the listening condition. In ideal listening conditions, very high levels of speech understanding can be achieved with a small number (e.g., four to six) of spectral channels (Shannon *et al.*, 1995; Dorman *et al.*, 1997). In more difficult listening situations (e.g., in the presence of multiple talkers or noise) performance increases with an increasing number of spectral channels up to at least 20 (Dorman *et al.*, 1997; Loizou *et al.*, 1999; Friesen *et al.*, 2001; Shannon *et al.*, 2004). While a similar trend has been observed in actual cochlear implant recipients, it appears most users receive relatively few effective channels of spectral information (i.e., four to ten) despite a larger number of available intracochlear electrodes (Holmes *et al.*, 1987; Fishman *et al.*, 1997; Friesen *et al.*, 2001). In the current study, the number of channels is manipulated to determine the differential impact of reverberation and masking as the amount of available spectral information varies.

In addition to studying the effects of reverberation in simulations of cochlear implant listening, the present investigation also considers differences between noise maskers and competing speech maskers in these same listening conditions. Recent papers by Qin and Oxenham (2003) and Stickney *et al.* (2004) demonstrate that under some implant simulation conditions, speech maskers are more efficient than noise maskers even though the noise maskers are more efficient when used with unprocessed speech. This finding has been interpreted as showing that the loss of temporal fine structure and pitch cues that results from cochlear implant processing leads to great difficulty in speech sound source segregation (Qin and Oxenham, 2003; Stickney *et al.*, 2004). In separate conditions, the current study uses continuous noise masking as well as masking from a combination of two talkers. The comparison between the efficiency of these two types of maskers will add new data relevant to the question of the role that informational masking plays in cochlear implant listening.

The present investigation consisted of two experiments. Experiment 1 investigated the impact of reverberation on speech understanding in quiet as a function of the number of available spectral channels. Experiment 2 examined the individual and combined effects of noise and reverberation with continuous noise and competing speech maskers in 6-channel and 12-channel simulations.

II. EXPERIMENT 1: EFFECT OF REVERBERATION IN QUIET

A. Method

1. Subjects

Sixteen students with normal hearing (audiometric thresholds $\leqslant$20 dB HL at 500, 1000, 2000, 4000, and 6000 Hz) from the University of Massachusetts Amherst participated as subjects. Subjects ranged in age from 19 to 29 years, were native speakers of English, and had no prior experience listening to vocoded speech. Partial course credit was awarded for participation in this experiment.

2. Speech materials

Stimuli consisted of lists of sentences (Helfer and Freyman, 2004) containing three key words, each of which were one- or two-syllable nouns or verbs from the Francis and Kucera (1982) list of most common words. The sentences were assigned to 1 of 23 topics (e.g., food, clothing, politics). In conditions where the target speaker was masked by other talkers, subjects were expected to need a prompt as to which speaker they should listen to. This is due to the fact that cochlear implant processing eliminates most fundamental frequency cues, thereby making different talkers sound similar. With the use of topic-based sentences, we were able to visually provide the subjects with the topic of the target sentence prior to its presentation. This sort of prompting would not have been possible with more commonly used speech corpuses such as the Hearing in Noise Test (HINT) sentences.

The topic sentences were spoken by a female talker with a standard American English dialect and digitally recorded in a sound-treated booth (IAC 1604) with 16-bit resolution at a 22 050-Hz sampling rate. Sentences were scaled to the same overall rms level. Recognition data acquired in speech-spectrum noise at a −13-dB signal-to-noise ratio (Helfer and Freyman, 2004) were used to group subsets of the sentences into 16 equally intelligible lists of 20 sentences with average unprocessed key word recognition of 50% correct.

3. Reverberation simulation

Figure 1 provides a flow-diagram of the signal processing utilized in this investigation, including the reverberation simulation described here. Using methods developed by Zurek *et al.* (2004), recordings for each list were processed under one of four reverberant conditions as follows. An

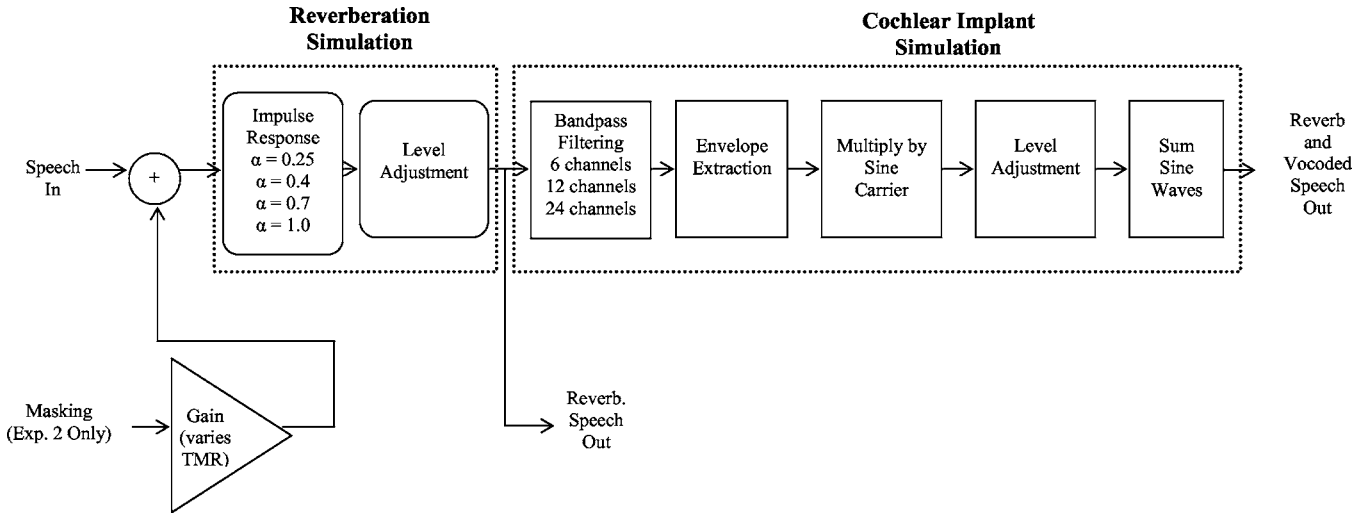


FIG. 1. The signal processing chain including both reverberation simulation and cochlear implant simulation. All stimuli entered the chain at the point of reverberation simulation. In experiment 1, stimuli processed for a subset of conditions exited the routine prior to cochlear implant simulation (reverb. speech out). All stimuli processed for experiment 2 were subjected to both reverberation and cochlear implant simulation (reverb+vocoded speech out).

image-method simulation of sound transmission (Allen and Berkley, 1979) was used to generate impulse responses from a source to a sensing point located on a model of a listener's head. This model was a rigid sphere, 8.75 cm in diameter, with a sensing point located 100° to the right along the sphere. The target source was located at 0°, 4 m from the sphere's center. The simulations modeled the acoustics of a relatively small rectangular classroom ($4.6 \times 6.6 \times 2.6$ m^3) at the University of Massachusetts Amherst. Figure 2 provides details of the dimensions of the room, the orientation of the sphere, and the relative positions of the sphere and the target source. For the purposes of the simulation, it was assumed that this room had uniform, frequency-independent absorption on all six surfaces. For further details of this simulation method, please see Zurek *et al.* (2004).

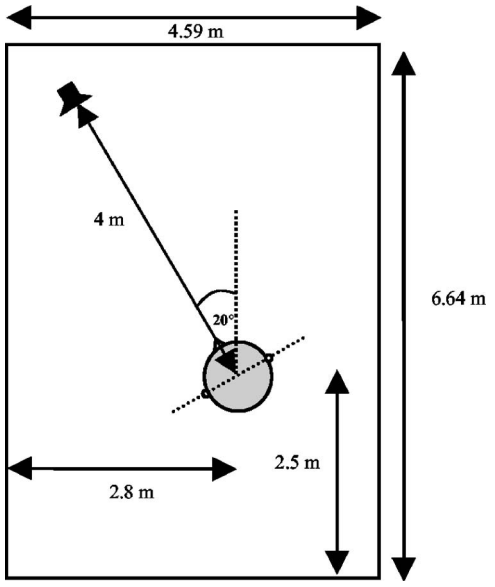


FIG. 2. Schematic illustrating the dimensions of the room, the orientation of the sphere, and the relative positions of the sphere and the target source within the room used in the reverberation simulation (adapted from Zurek *et al.*, 2004).

The current simulations varied the absorption coefficient value (α) from 0.25, to 0.4, to 0.7, to 1.0, which resulted in estimated reverberation times (T) of 0.425, 0.266, 0.152, and 0.0 s, respectively (Zurek *et al.*, 2004). The impulse responses incorporated room reflections for a period of 0.5 s following the initial direct impulse response (11 025 points at a 22 050 sampling rate). Note that the length of the impulse response was longer than the longest reverberation time simulated. So while at longer simulated reverberation times decay may not have reached 0 by the end of the impulse response, the signal was reduced in level by greater than 60 dB by this point in time, resulting in an extremely small amount of truncation of only very low-level sound energy. Once reverberation was simulated by convolving each sentence with the room impulse response under test, the resulting sentence was then scaled to a rms level of 65 dBA to adjust for the increase in power caused by reverberant sound. While this level adjustment may have been somewhat artificial, our primary interest in this experiment was in the temporal smearing aspect of reverberation.

4. Cochlear implant simulation

As illustrated in Fig. 1, the reverberated sentences were either processed by several different tone-excited channel vocoder systems (i.e., systems with 6, 12, and 24 channels) implemented in Matlab, or subjected to no further processing or reduction in bandwidth (subsequently referred to as unprocessed). The implementations of the vocoder systems followed those of Qin and Oxenham (2003). Unprocessed speech was filtered (sixth-order Butterworth) into contiguous frequency bands in the 80–6000 Hz range, each with equal bandwidth as measured on an ERB scale (Glasberg and Moore, 1990). The center frequencies and bandwidths for each of the processors are provided in Table I. Envelopes for each band were obtained from filtered speech by half-wave rectification and smoothing with a second-order Butterworth low-pass filter. The bandwidth of the smoothing filter was the smaller of 300 Hz or half the analysis bandwidth. The result-

TABLE I. Center frequencies and bandwidths, expressed in Hz, for each of three experimental vocoder systems.

6 channels	Center frequency	180	446	885	1609	2803	4773					
	Bandwidth	201	331	546	901	1487	2453					
12 channels	Center frequency	124	224	353	519	731	1005	1355	1806	2385	3128	4084	5310
	Bandwidth	88	113	145	186	239	307	395	507	651	836	1074	1379
24 channels	Center frequency (1–12)	101	145	194	251	315	387	469	562	668	787	923	1077
	Bandwidth (1–12)	41	47	53	60	68	77	87	99	112	127	144	163
	Center frequency (13–24)	1251	1448	1671	1925	2212	2538	2906	3324	3798	4335	4944	5634
	Bandwidth (13–24)	185	210	238	269	305	346	392	444	503	571	647	733

ing envelope in each band was used to amplitude modulate a sine wave at the band's center frequency. The modulated carriers were level matched to their original in-band input signals and summed across channels to produce simulated cochlear implant-processed speech.

The decision to replace the noise-band carriers used by Qin and Oxenham (2003) and other investigators (Shannon *et al.*, 1995; Fu *et al.*, 1998; Faulkner *et al.*, 2000; Friesen *et al.*, 2001; Green *et al.*, 2002; Stickney *et al.*, 2004) with tone carriers requires some discussion. Dorman *et al.* (1997) compared tone-excited vocoders with noise-excited vocoders after observing that many cochlear implant users perceive individual channel stimulation as tonelike signals rather than noiselike signals. The results of their simulations using both tone-excited and noise-excited vocoders with *experienced* listeners suggest that the two vocoders provide similar levels of performance. Two later experiments conducted in our lab (Whitmal *et al.*, 2004, 2005) indicate substantial differences between tone-vocoded speech and noise-vocoded speech for *inexperienced* listeners. A similar phenomenon was also observed by Dorman *et al.* (1997). Likewise, in pilot tests for an experiment on speaker gender discrimination with cochlear implants, Fu *et al.* (2004) found that performance with noise-excited vocoders was at the level of chance (far below that achieved by actual cochlear implant users) while performance with tone-excited vocoding was significantly higher and better at predicting scores of the highest performing cochlear implant users. Similar results were more recently obtained in an experiment on gender identification (Gonzalez and Oliver, 2005). Considering these results, it is reasonable to expect that tone-excited vocoding will provide performance that is better and perhaps more characteristic of cochlear implant performance than noise-excited vocoding.

5. Procedures

Signal calibration was completed daily by playing a recording of speech spectrum noise (level matched in long-term spectrum and rms level to the target stimuli), convolved with the impulse response of the anechoic room through the same hardware used in the experiment, and adjusting the gain electronically until the target level of 65 dBA was achieved. For data collection, subjects were seated in a double-walled sound-treated booth (IAC 1604) where a laptop screen prompted them with the word "Ready?" and the sentence topic 2 s before the sentence was presented. The sentence was delivered from a computer's sound card (Analog Devices, SoundMax Integrated Digital Audio) using 16-bit resolution at 22 050 Hz, passed through an attenuator (TDT PA4) and headphone amplifier (TDT HB5) and presented to the subjects' right ears (corresponding to right sensing point on the sphere used to simulate the head) over a TDH-50P headphone. Subjects repeated each sentence they heard into a talk-back microphone monitored by a researcher, who recorded the number of key words correctly repeated. Practice was provided and consisted of the presentation of five sentences in each of eight conditions, which were combinations of various vocoders (6, 12, 24, and unprocessed) and absorption coefficients (α=0.4, and 1.0). No feedback was given during the practice session. Sentences used for practice were not used in the main experiment.

In the main experiment, one combination of the four reverberation conditions (α=0.25, 0.4, 0.7, and 1.0) and four cochlear implant simulation conditions (6, 12, 24, and unprocessed) was used to present one list of sentences in each of 16 different experimental conditions. Condition/list pairings for the subjects and presentation order of the conditions was determined by a 16×16 Latin square. This design resulted in all lists being presented in all positions one time, thereby controlling for any list effects. Subjects typically completed data collection during one 2-h session with breaks provided as needed.

B. Results

Figure 3 presents the results for the four vocoded conditions as a function of the absorption coefficient employed in the reverberation simulation. Results indicated that speech

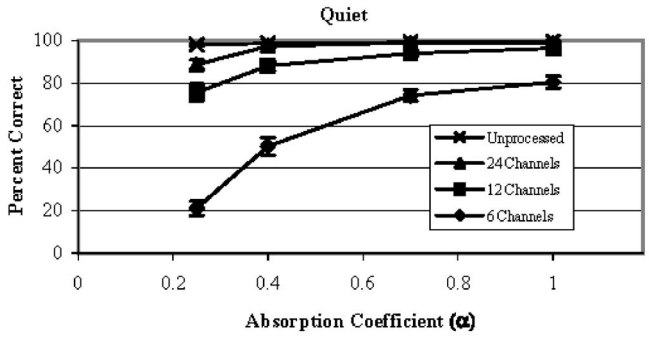


FIG. 3. Speech recognition performance for the 6-, 12-, and 24-channel vocoders and for unprocessed/natural speech as a function of absorption coefficient (α). Error bars represent ±1 standard error.

perception remained excellent in the presence of varying levels of reverberation for the unprocessed and the 24-channel conditions. There was little difference between the 12- and 24-channel vocoders with absorption coefficients as small as 0.4. However, with an absorption coefficient of 0.25, performance was notably poorer with the 12-channel system. Speech understanding with the 6-channel processor was significantly more susceptible to the impact of reverberation for all absorption coefficient levels.

Average results with the 6-channel processor in the anechoic condition ($\alpha=1.0$) were 81%, which is somewhat lower than those obtained in our laboratory by the same 6-channel system without reverberation simulation where levels ranged from 89% to 92% (Whitmal *et al.*, 2004, 2005). While different subjects participated in each of these experiments, we attribute this difference primarily to the fact that subjects in the current investigation listened monaurally while subjects in the previous investigations listened diotically.

A repeated measures analysis of variance was conducted using intelligibility scores converted to rationalized arcsine units (Studebaker, 1985). Results of the analysis demonstrated a significant main effect for absorption coefficient ($F=230.99$, $p<0.0001$) and for the number of channels ($F=855.53$, $p<0.0001$) as well as a significant interaction between absorption coefficient and number of channels ($F=31.03$, $p<0.00001$). The nature of the interaction effect can be seen in Fig. 3, which illustrates that conditions employing fewer channels of spectral information are more significantly affected by higher levels of reverberation (lower absorption coefficients). *Posthoc* Scheffe testing indicated performance was significantly different and poorer as the absorption coefficient decreased from 1.0, to 0.7, to 0.4, and to 0.25. Performance became significantly poorer as the stimuli went from the unprocessed condition to the 24-channel vocoder, to the 12-channel vocoder, and finally to the 6-channel vocoder system.

III. EXPERIMENT II: COMBINED EFFECT OF REVERBERATION AND MASKING

A. Method

1. Subjects

Ten students with normal hearing (audiometric thresholds ≤20 dB HL at 500, 1000, 2000, 4000, and 6000 Hz) from the University of Massachusetts Amherst participated as subjects. Subjects ranged in age from 18 to 22 years, were native speakers of English, and had no prior experience listening to vocoded speech. Partial course credit was awarded for participation in this experiment.

2. Speech materials

Speech materials and recordings were identical to those described in experiment 1, except that 480 sentence recordings were divided into 30 equally intelligible lists of 16 sentences.

3. Addition of maskers

Preprocessing for 24 of the 30 lists consisted of electronically adding either speech-spectrum noise (SSN) or two-talker babble (TTB) at a +8 or +18 dB target-to-masker ratio (TMR) prior to reverberation and cochlear implant simulation. These TMRs were chosen based upon previous work in our lab which showed that they are sensitive enough to show effects of parameter manipulation (e.g., channel number) while avoiding ceiling and/or floor effects (Whitmal *et al.*, 2004, 2005).

In order to generate the SSN, a white noise signal (10 s long, sampled at a rate of 22 050 Hz) was passed through an all-pole filter (51 coefficients) derived from a linear prediction model of the average power spectrum of the individual target sentences. The resultant SSN was scaled to the rms level of the original sentences.

The two-talker babble (used in Freyman *et al.*, 2001) consisted of digital recordings (sampling rate: 22 050 Hz) of two college-aged female students speaking different nonsense sentences. Pauses between sentences were removed to produce two recordings of continuous speech that were matched in rms level. A 35-s-long stream of the babble was extracted for use in the present investigation.

Maskers for each sentence consisted of segments of either the 10-s SSN signal or the 35-s TTB signal described above. The segments were the same length as the individual sentence recordings. Each sentence recording contained a short silent period of variable length both before (estimated mean=0.064 s) and after (estimated mean=0.328 s) the sentence was spoken. The offsets of the masker segments were selected at random to prevent the subjects from hearing the same phrases or sections repeatedly. The masker segment was scaled to produce the desired TMR for the particular trial, and the target and masker were summed together. The composite signal was then filtered by a second-order Butterworth band-pass filter with cutoff frequencies of 80 and 6000 Hz.

4. Reverberation simulation

Reverberation was simulated as in experiment 1. Based upon the results of experiment 1, absorption coefficients chosen for this experiment included 1.0, 0.7, and 0.4. The $\alpha=0.25$ condition from experiment 1 was eliminated as we felt performance would be at floor levels when noise and reverberation were combined.

5. Cochlear implant simulation

Following reverberation simulation, the sentence lists were then processed using either 6- or 12-channel vocoding as described in experiment 1. These numbers of channels were chosen as they are most closely representative of the numbers of perceptual channels actual cochlear implant users are estimated to receive.

6. Procedures

Daily calibration was accomplished as in experiment 1. Practice consisted of the presentation of five sentences in 18 different conditions which were randomized according to a

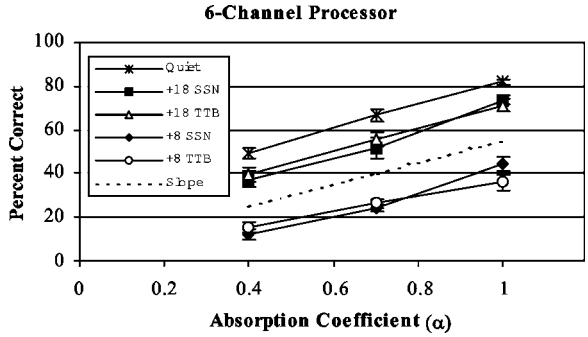


FIG. 4. Speech recognition performance for the 6-channel vocoder in quiet and in both speech spectrum noise and two-talker babble at TMRs of +18 and +8 as a function of absorption coefficient (α). Error bars represent ±1 standard error.

Latin square design. The 18 conditions included combinations of the following: maskers/TMRs (quiet, +8 SSN, +8 TTB), channels (6 channels, 12 channels), and absorption coefficients (α=0.4, 0.7, 1.0). No feedback was given during the practice session. Subjects typically completed data collection in two 90-min sessions held on separate days. All other procedures were identical to those described in experiment 1.

B. Results

Figures 4 and 5 present the results for various types and levels of masking as a function of absorption coefficient for the 6- and 12-channel vocoders, respectively. Results illustrate that the combination of noise and reverberation was extremely detrimental to speech perception. In the most difficult experimental conditions (6-channel processor, +8 dB TMR, α=0.4), subjects understood only 12% (in SSN) and 15% (in TTB) of the key words. Based upon the results obtained by Helfer and Freyman (2005), we predicted that listeners with normal hearing would demonstrate nearly perfect performance when listening to unprocessed/natural speech under similar conditions.

A repeated measures analysis of variance was conducted using intelligibility scores converted to rationalized arcsine units (Studebaker, 1985). Results of the analysis demonstrated significant main effects for absorption coefficient (F=190.98, $p<0.0001$), the number of channels (F=789.06, $p<0.0001$), and for noise (F=143.11, $p<0.0001$). There were no significant interaction effects. This is illustrated by the fact that the slopes of the lines in Figs. 4 and 5 are nearly identical. *Posthoc* tests according to Scheffe indicated that performance decreased significantly as the absorption coefficient decreased from 1.0 to 0.7 to 0.4 and as the number of channels decreased from 12 to 6. Performance in quiet was significantly better than performance in any of the noise conditions. Further, there was a significant difference between TMRs (i.e., +18 and +8) for both types of noise (i.e., SSN and TTB), but not between noise types for either TMR.

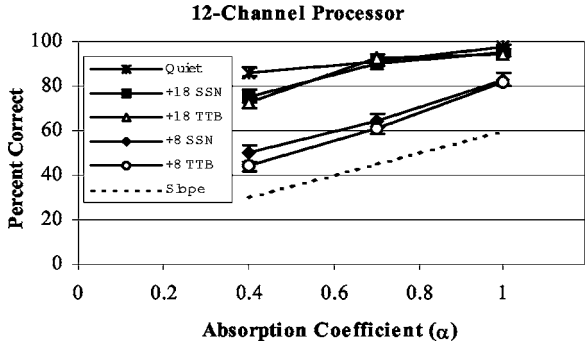


FIG. 5. Speech recognition performance for the 12-channel vocoder in quiet and in both speech spectrum noise and two-talker babble at TMRs of +18 and +8 as a function of absorption coefficient (α). Error bars represent ±1 standard error.

IV. DISCUSSION

A. Reverberation in quiet

Listening through the limited number of spectral channels available to a cochlear implant user may be thought of as a form of spectral smearing analogous to the broadened peripheral analysis filters characteristic of acoustic listeners with hearing loss (Shannon and Fu, 1999; Shannon *et al.*, 2004). Despite such "smearing," high levels of speech understanding in quiet are possible provided a relatively pristine (in the time domain) speech signal is available to the cochlear implant user. When reverberation is present, the temporal envelope of the speech signal itself is smeared prior to its processing by the cochlear implant system and the recipient's altered auditory system. Therefore, reverberation results in a corruption of the primary cues cochlear implant users are reliant on for speech understanding. In the current simulations in quiet, the 12-channel and especially the 24-channel processing systems were shown to be largely resistant to the tested levels of reverberation. However, the results for the 6-channel system showed substantial degradation of speech understanding as reverberation increased. Six channels is within the range of effective spectral channels available to actual cochlear implant users (Holmes *et al.*, 1987; Fishman *et al.*, 1997; Friesen *et al.*, 2001). Thus, to the extent that the simulations are good models of actual cochlear implant listening, the results suggest that for an average cochlear implant recipient, even small amounts of reverberation will be extremely detrimental to speech understanding.

It is important to note that none of the absorption coefficients used in these experiments represent an extreme amount of reverberation. In fact, the lowest absorption coefficient (α=0.25, $T\approx$0.425 s) represents a considerably more favorable listening condition than that of the actual classroom used as a model in the simulation, which had an estimated average absorption coefficient of 0.14 (Zurek *et al.*, 2004). According to Beranek's (1954) classification system, the present simulation with α=0.25 would be considered to be "medium dead," and the room with α=0.4 would be considered "dead." Cochlear implant users certainly must converse in environments that have significantly more reverberation than those simulated here.

B. Reverberation combined with masking

When the target speech signal was combined with competing speech or noise prior to processing with the reverberant impulse response, the resulting speech recognition scores

were alarmingly poor. As with the reverberant conditions, the masking conditions used in this investigation (TMR's of +8 and +18 dB) would be considered to be quite favorable for listeners with normal hearing. For example, with these same sentence stimuli, subjects in Helfer and Freyman (2005) obtained 85% correct key-word recognition at 0 dB TMR, which is 8 dB lower than in the poorest TMR condition in the current study. For continuous discourse with noise masking, Plomp (1976) found that with a reverberation time of 0.4 (similar to that estimated for our $\alpha=0.25$) and monaural listening (as in the current study), listeners obtained "intelligibility thresholds" at a TMR of 0 dB. In the current study, with a more favorable reverberation time ($\alpha=0.4$, $T=0.266$), and a much more positive TMR of +8 dB in speech-spectrum noise, listeners recognized only 12% of key words through the 6-channel system. Clearly, the 6-channel processing greatly increases listeners' susceptibility to reverberation and noise.

For a given TMR condition (+8 dB, +18 dB, quiet), Fig. 4 shows that for the 6-channel system, speech recognition performance declined linearly as the absorption coefficient decreased. The approximate slope, shown by the dashed line in the figure, is 5 percentage points per 0.1 change in α. The same slope describes the 12-channel data in Fig. 5, at least for the +8 dB TMR. The pattern of data in the +18 dB TMR and quiet conditions appears to show a slight ceiling effect. The linear relationships allow for straightforward comparison of the magnitude of the effects of noise and reverberation. It appears from the figures that increasing reverberation from an α of 1.0 to an α of 0.4 has approximately the same (30 percentage point) effect as decreasing the TMR by 10 dB from +18 to +8 dB. This reveals a surprisingly large effect of reverberation, especially considering that the reverberation times created were not extreme.

The relatively parallel data obtained for different TMRs indicate additivity of the effects of noise and reverberation. For example, for the 6-channel system, speech-spectrum noise at a +8 dB TMR creates a loss of 38 percentage points from the quiet nonreverberant condition (82% versus 44%). Changing reverberation in quiet from fully absorbent surfaces ($\alpha=1.0$) to an absorption coefficient of 0.4 creates a loss of 33 percentage points (82% versus 49%). Introducing both noise and reverberation to the degrees just described creates a loss of 70 percentage points (82% versus 12%), only one point less than the simple addition of the individual effects (71 percentage points).

The simple additivity of noise and reverberation effects is different from some of the findings obtained with unprocessed/natural speech in normal-hearing and hearing-impaired listeners. Interactions between the effects of noise and reverberation have been reported (Nabelek and Mason, 1981; Loven and Collins, 1988; Helfer and Wilber, 1990; Payton *et al.*, 1994), although some of these studies do show additivity over certain ranges of noise and reverberation.

C. STI analysis

The lack of an interaction between masking and reverberation in our simulation of cochlear implant processed speech may be related to the listener's heavy reliance on temporal envelope cues that are degraded by both noise and reverberation. Both distortions are known to affect the modulation spectrum, and therefore the speech transmission index (STI), a single number value derived from the modulation spectrum across frequency channels (Houtgast and Steeneken, 1971; Steeneken and Houtgast, 1980). Previous work has shown that the STI correlates strongly with the effects of noise and reverberation on speech intelligibility (e.g., Duquesnoy and Plomp, 1980; Houtgast *et al.*, 1980; Payton *et al.*, 1994). To investigate how well the combined effects of masking, reverberation, and our implant simulation were captured by the speech transmission index, we performed an STI analysis on our stimuli and evaluated its predictive value for the speech recognition results.

The STI is a frequency-weighted average of octave-band apparent signal-to-noise ratios (aSNRs), each of which may range from −15 dB (representing poor intelligibility) to +15 dB (representing excellent intelligibility). The STI is then calculated via the relation

$$\mathrm{STI}=\sum_{i=1}^{7} w_i\left[\frac{\mathrm{aSNR}_i+15}{30}\right], \tag{1}$$

where i denotes the number of the octave band (125, 250, 500, 1000, 2000, 4000, or 8000 Hz, respectively), aSNR_i denotes the aSNR for band i, and w_i is an empirically derived weight for band i (Steeneken and Houtgast, 1980). The allowable range of aSNR values produces a corresponding range of STI values from 0 (poor intelligibility) to 1 (excellent intelligibility). aSNR values, in turn, are calculated from modulation transfer functions (MTFs) that quantify changes in modulation depth:

$$\mathrm{aSNR}_i=10\log_{10}\left(\frac{\mathrm{MTF}_i}{1-\mathrm{MTF}_i}\right). \tag{2}$$

The MTFs for each band range are traditionally calculated from averages of 1/3-octave-band envelope power spectra over a range of 0.63 to 12.5 Hz, normalized by average envelope energy.

It is important to note that the STI has been less useful for predicting and explaining intelligibility changes caused by nonlinear processing such as envelope thresholding (Drullman, 1995) and spectral subtraction (Ludvigsen *et al.*, 1993). One might expect similar difficulties with vocoder-processed speech, which relies heavily on nonlinear processing. Hence, the STI calculations presented here rely on the work of Goldsworthy and Greenberg (2004), who attribute prediction errors to flaws in intermediate MTF calculations, and proposed simple and effective modifications to address them. This work uses their modified version of the envelope regression method (Ludvigsen *et al.*, 1990), which models the input-output envelope transformation as

$$MTF_i=\beta\frac{\lambda_{xy,i}}{\lambda_{x,i}}, \tag{3}$$

with $\lambda_{\mathrm{xy,i}}$ being the respective cross correlation of the discrete-time input envelope $x_i[n]$ and output envelope $y_i[n]$ in band i, $\lambda_{x,i}$ being the variance of $x_i[n]$, and β being

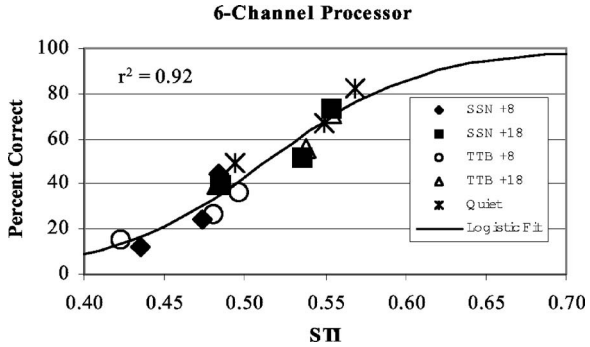


FIG. 6. Speech recognition performance for the 6-channel vocoder as a function of speech transmission index (STI). Each symbol represents one masker condition across different levels of absorption. The solid line represents the predictive equation for the current data.

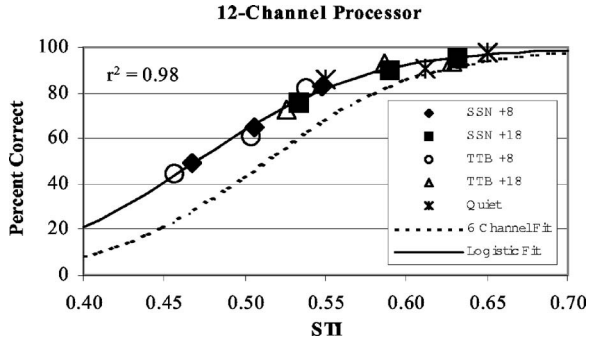


FIG. 7. Speech recognition performance for the 12-channel vocoder as a function of speech transmission index (STI). Each symbol represents one masker condition across different levels of absorption. The solid line represents the predictive equation for the current data. For comparative purposes, the predictive equation for the 6-channel system is presented as a dashed line.

a normalization coefficient. For cases of noise and/or reverberation, β simply equals $\mu_{x,i}/\mu_{y,i}$, the time average of $x_i[n]$ divided by the time average of $y_i[n]$. The envelope regression method is chosen for two reasons. First, its results correlate well with those of traditional STI methods in noise and reverberation. Second, its computational efficiency is better than those of the other proposed methods.

STI values for all combinations of vocoder, noise, and reverberation processing were computed in Matlab on a personal computer, using a database of 15 digitally recorded sentences from experiments 1 and 2. Each sentence (in both unprocessed and processed form) was passed through a bank of seven eighth-order Butterworth octave-band filters centered at frequencies ranging from 125 to 8000 Hz. The output of each bandpass filter was then squared, smoothed by a 200-Hz fourth-order Butterworth low-pass filter, and downsampled to 551.25 Hz to produce band-limited envelope functions $x_i[n]$ (for unprocessed speech) and $y_i[n]$ (for processed speech). Values of $\lambda_{xy,i}$ and $\lambda_{x,i}$ were computed as

$$\lambda_{xy,i} = \frac{1}{N-1}\sum_{n=1}^{N}(x_i[n]-\mu_{x,i})(y_i[n]-\mu_{y,i}) \quad (4)$$

and

$$\lambda_{x,i} = \frac{1}{N-1}\sum_{n=1}^{N}(x_i[n]-\mu_{x,i})^2, \quad (5)$$

where N is the length (in samples) of $x_i[n]$ and $y_i[n]$. These values were then used to compute values of MTF_i for the processed sentence. STI values were then calculated in accordance with Eqs. (1) and (2).

As can be seen in Figs. 6 and 7, a reasonably good prediction of performance from the STI was obtained for both the 6- and 12-channel vocoders. However, the predictor functions for 6-and 12- channel vocoders are substantially different from each other. This difference appears to be due to the effect of vocoding on STI computation. Preliminary results comparing STI values computed under identical environmental conditions with and without vocoding show that vocoded STI values are compressed relative to their nonvocoded counterparts, and approach an asymptotic maximum which is dependent on the number of channels. Hence, nonvocoded STI values ranging from 0.6 to 1.0 correspond to average STI values of (approximately) 0.6 for vocoded systems with eight or more channels.

D. Effect of masker type

As shown in Figs. 4 and 5, the choice of type of masker (noise or two-talker competing speech) did not affect speech results appreciably. This can be contrasted with data from Qin and Oxenham (2003) and Stickney *et al.* (2004), who found that under some vocoding conditions, single-talker competing speech was significantly more effective than noise masking. The findings in these papers are especially interesting because when the speech was unprocessed, the reverse was true, i.e., the noise masker was more effective than the competing speech masker. The difference was interpreted as indicating that the vocoding process makes it difficult to distinguish between target and masking voices, possibly leading to an increase in informational masking (Stickney *et al.*, 2004). This hypothesis is bolstered by the fact that neither Qin and Oxenham nor Stickney *et al.* found that single talkers of the same gender as the target talker produced more masking than single talkers of a different gender, a result observed in normal-hearing individuals listening to unprocessed speech and attributed to informational masking (Brungart, 2001). It is possible that lack of sufficient access to fundamental frequency cues through cochlear implant processing leads to the elimination of the gender-specific differences in masking observed in normally hearing listeners.

Our study differs from Stickney *et al.* and Qin and Oxenham in two ways: we tested in both anechoic and reverberant conditions and we employed two talkers as our maskers as opposed to one. As we observed similarities in performance between SSN and TTB in both anechoic and reverberant conditions it is most likely that these observations relate directly to the difference in the number of masking talkers. With unprocessed speech, previous investigations (e.g., Miller, 1947; Yost *et al.*, 1996, Brungart *et al.*, 2001; Freyman *et al.*, 2004) have shown that two talkers are more effective maskers than are single talkers. In particular Brungart *et al.* (2001) and Freyman *et al.* (2004) suggest that two-talker maskers produce more informational masking

than single talkers. If this same difference occurred with cochlear implant processed speech, we would have expected an effect in the same direction of Stickney *et al.* (2004) and Qin and Oxenham (2003), but of even greater magnitude. The fact that we did not find any appreciable differences between SSN and TTB suggests that simulations of cochlear implant processing change the effect of number of competing talkers on the amount of informational and/or energetic masking produced.

V. CONCLUSIONS

In the presence of a limited number of spectral channels of information, low to moderate levels of reverberation result in decreased speech understanding relative to that demonstrated in nonreverberant conditions. While masking further decreases performance in such situations, its effect is simply additive. These results suggest that the speech understanding ability of an average cochlear implant user (e.g., one who receives 6 channels of spectral information) is likely to be severely compromised in real-world listening environments and that performance levels obtained in laboratories appear to overestimate these listeners' everyday functioning. Ongoing efforts to increase the number of effective channels available to cochlear implant recipients through the use of current steering to create additional "virtual" channels in present-day cochlear implant systems (Bonham *et al.*, 2005; Downing *et al.*, 2005) and feasibility studies of "optical cochlear implants" (Richter *et al.*, 2005) which may be used in future systems are clearly justified. Until the time this goal is realized, or even partially so, alternative means of improving cochlear implant users' function under realistic acoustic conditions must be considered. Results of the current investigation indicate one of the most effective ways of doing so would be to reduce the amount of reverberation present in the listening environment.

ACKNOWLEDGMENTS

This work was partially supported by the National Institutes of Health (NIDCD Grant Nos. R03 DC 7969-01 and R01 DC 01625). The authors thank Christine Alexander, Heather Nunes, Katelyn McLaughlin, and Beth Ann Jacques for their assistance in data collection.

Allen, J. B., and Berkley, D. A. (**1979**). "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**, 943–951.

Beranek, L. L. (**1954**). *Acoustics* (McGraw-Hill, New York).

Bonham, B., Snyder, R., Corbett, S., Johnson, T., and Rebscher, S. (**2005**). "Physiolgoical measures of auditory nerve activation by current steering," ARO Abstracts **28**, 359.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

Dorman, M. F., and Loizou, P. C. (**1998**). "The identification of consonants and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine channels," Ear Hear. **19**, 162–166.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.

Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (**1998**). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels," J. Acoust. Soc. Am. **104**, 3583–3585.

Downing, M., Firszt, J. B., Runge-Samuelson, C. L., Koch, D. B., and Litvak, L. (**2005**). "Current steering and perceptual channels in HiResolution bionic ear users: Multi-center study of cochlear-implant place/pitch relationship," ARO Abstracts **28**, 85–86.

Dreschler, W. A., and Leeuw, A. R. (**1990**). "Speech reception in reverberation related to temporal resolution," J. Speech Hear. Res. **33**, 181–187.

Drullman, R. (**1995**). "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**, 585–592.

Duquesnoy, A. J., and Plomp, R. (**1980**). "Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis," J. Acoust. Soc. Am. **68**, 537–544.

Faulkner, A., Rosen, S., and Smith, C. (**2000**). "Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: implications for cochlear implants," J. Acoust. Soc. Am. **108**, 1877–1887.

Faulkner, A., Rosen, S., and Stanton, D. (**2003**). "Simulations of tonotopically mapped speech processors for cochlear implant electrodes varying in insertion depth," J. Acoust. Soc. Am. **113**, 1073–1080.

Firszt, J. B., Holden, L. K., Skinner, M. W., Tobey, E. A., Peterson, A., Gaggl, W., Runge-Samuelson, C. L., and Wackym, P. A. (**2004**). "Recognition of speech presented at soft to loud levels by adult cochlear implant recipients of three cochlear implant systems," Ear Hear. **25**, 375–387.

Fishman, K. E., Shannon, R. V., and Slattery, W. H. (**1997**). "Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor," J. Speech Lang. Hear. Res. **40**, 1201–1215.

Francis, W. N., and Kucera, H. (**1982**). *Frequency Analysis of English Usage: Lexicon and Grammar* (Houghton Mifflin, Boston).

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.

Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.

Fu, Q. J., and Galvin, J. J., III (**2001**). "Recognition of spectrally asynchronous speech by normal-hearing listeners and Nucleus-22 cochlear implant users," J. Acoust. Soc. Am. **109**, 1166–1172.

Fu, Q. J., Chinchilla, S., and Galvin, J. J. (**2004**). "The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users," J. Assoc. Res. Otolaryngol. **5**, 253–260.

Fu, Q. J., Shannon, R. V., and Wang, X. (**1998**). "Effects of noise and spectral resolution on vowel and consonant recognition: acoustic and electric hearing," J. Acoust. Soc. Am. **104**, 3586–3596.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.

Goldsworthy, R. L., and Greenberg, J. E. (**2004**). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**, 3679–3689.

Gonzalez, J., and Oliver, J. C. (**2005**). "Gender and speaker identification as a function of the number of channels in spectrally reduced speech," J. Acoust. Soc. Am. **118**, 461–470.

Green, T., Faulkner, A., and Rosen, S. (**2002**). "Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants," J. Acoust. Soc. Am. **112**, 2155–2164.

Helfer, K. S., and Freyman, R. L. (**2004**). "Development of a topic-related sentence corpus for speech perception research," J. Acoust. Soc. Am. **115**, 2601.

Helfer, K. S., and Freyman, R. L. (**2005**). "The role of visual speech cues in reducing energetic and informational masking," J. Acoust. Soc. Am. **117**, 842–849.

Helfer, K. S., and Wilber, L. A. (**1990**). "Hearing loss, aging, and speech perception in reverberation and noise," J. Speech Hear. Res. **33**, 149–155.

Holmes, A., Kemker, F. J., and Merwin, G. (**1987**). "The effects of varying the number of cochlear implant electrodes on speech perception," Am. J.

Otol. **8**, 240–246.
Houtgast, T., and Steeneken, H. J. M. (**1971**). "Evaluation of speech transmission channels by using artificial signals," Acustica **25**, 355–367.
Houtgast, T., Steeneken, H. J. M., and Plomp, R. (**1980**). "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," Acustica **46**, 60–72.
Irwin, R. J., and McAuley, S. F. (**1987**). "Relations among temporal acuity, hearing loss, and the perception of speech distorted by noise and reverberation," J. Acoust. Soc. Am. **81**, 1557–1565.
Koch, D. B., Osberger, M. J., Segel, P., and Kessler, D. (**2004**). " HiResolution™ and conventional sound processing in the HiResolution™ bionic ear: Using appropriate outcome measures to assess speech recognition ability," Audiol. Neuro-Otol. **9**, 214–223.
Lochner, J. P. A., and Burger, J. F. (**1961**). "The intelligibility of speech under reverberant conditions," Acustica **11**, 195–200.
Lochner, J. P. A., and Burger, J. F. (**1964**). "The influence of reflections on auditorium acoustics," J. Sound Vib. **1**, 426–454.
Loizou, P. C., Dorman, M., and Tu, Z. (**1999**). "On the number of channels needed to understand speech," J. Acoust. Soc. Am. **106**, 2097–2103.
Loizou, P. C., Dorman, M., Poroy, O., and Spahr, T. (**2000**). "Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution," J. Acoust. Soc. Am. **108**, 2377–2387.
Loven, F. C., and Collins, M. J. (**1988**). "Reverberation, masking, filtering, and level effects on speech recognition performance," J. Speech Hear. Res. **31**, 681–695.
Ludvigsen, C., Elberling, C., and Keidser, G. (**1993**). "Evaluation of a noise reduction methods—Comparison of observed scores and scores predicted from STI," Scand. Audiol. Suppl. **38**, 50–55.
Ludvigsen, C., Elberling, C., Deidser, G., and Poulsen, T. (**1990**). "Prediction of intelligibility of non-lineraly processed speech," Acta Oto-Laryngol., Suppl. **469**, 190–195.
Miller, G. A. (**1947**). "The masking of speech," Psychol. Bull. **44**, 105–129.
Muller, J., Schon, F., and Helms, J. (**2002**). "Speech understanding in quiet and noise in bilateral users of the MED-EL COMBI 40/40+ cochlear implant system," Ear Hear. **23**, 198–206.
Nabelek, A. K., and Mason, D. (**1981**). "Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms," J. Speech Hear. Res. **24**, 375–383.
Nabalek, A. K. N., and Robinette, L. (**1978**). "Influence of the precedence effect on word identification by normally hearing and hearing impaired subjects," J. Acoust. Soc. Am. **63**, 187–194.
Payton, K. L., Uchanski, R. M., and Braida, L. D. (**1994**). "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," J. Acoust. Soc. Am. **95**, 1581–1592.
Plomp, R. (**1976**). "Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source," Acustica **34**, 200–210.
Qin, M. K., and Oxenham, A. J. (**2003**). "Effects of simulated cochlear-implant-processing on speech reception in fluctuating maskers," J. Acoust. Soc. Am. **114**, 446–454.
Richter, C -P., Izzo, A., Walsh, J. T., and Jansen, E. D. (**2005**). "Optically-evoked acoustic nerve activity," ARO Abstracts **28**, 354.
Shannon, R. V. (**1989**). "Detection of gaps in sinusoids and pulse trains by patients with cochlear implants," J. Acoust. Soc. Am. **85**, 2587–2592.
Shannon, R. V. (**1993**). "Quantitative comparison of electrically and acoustically evoked auditory perception: implications for the location of perceptual mechanisms," Prog. Brain Res. **97**, 261–269.
Shannon, R. V., and Fu, Q. J. (**1999**). "Contributions of spectral resolution to speech recognition in noise," J. Acoust. Soc. Am. **105**, 1238.
Shannon, R. V., Fu, Q-J., and Galvin, J., III (**2004**). "The number of spectral channels required for speech recognition depend on the difficulty of the listening situation," Acta Oto-Laryngol., Suppl. **552**, 50–54.
Shannon, R. V., Zeng, F-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Skinner, M. W., Arndt, P. L., and Staller, S. J. (**2002**). "Nucleus ® 24 advanced encoder conversion study: Performance versus preference," Ear Hear. **23**, 2S–17S.
Spahr, A. J., and Dorman, M. F. (**2004**). "Performance of subjects fit with Advanced Bionics CII and Nucleus 3G cochlear implant devices," Arch. Otolaryngol. Head Neck Surg. **130**, 624–628.
Steeneken, H. J. M., and Houtgast, T. (**1980**). "A physical method for measuring speech transmission quality," J. Acoust. Soc. Am. **67**, 318–326.
Stickney, G. S., Zeng, F-G., Litovsky, R., and Assmann, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. **116**, 1081–1091.
Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transformation," J. Speech Hear. Res. **28**, 455–462.
Whitmal, N. A., III, Poissant, S. F., Freyman, R. L., and Helfer, K. S. (**2004**). "Effect of combining different carriers across bands on speech intelligibility in cochlear implantsimulation," ARO Abstracts **27**, 29.
Whitmal, N. A., III, Poissant, S. F., Freyman, R. L., and Helfer, K. S. (**2005**). "Effect of low-noise on speech intelligibility in simulations of cochlear implant processors with normal-hearing listeners," ARO Abstracts **28**, 260.
Yost, W. A., Dye, R. H., and Sheft, S. (**1996**). "A simulated cocktail party with up to three sound sources," Percept. Psychophys. **58**, 1026–1036.
Zeng, F. G., and Galvin, J. J., III (**1999**). "Amplitude mapping and phoneme recognition in cochlear implant listeners," Ear Hear. **20**, 60–74.
Zurek, P. M., Freyman, R. L., and Balakrishnan, U. (**2004**). "Auditory target detection in reverberation," J. Acoust. Soc. Am. **115**, 1609–1620.

A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners[a)]

Sygal Amitay,[b)] Amy Irwin, David J. C. Hawkey, Justin A. Cowan, and David R. Moore
MRC Institute of Hearing Research, University Park, Nottingham, NG7 2RD, United Kingdom

(Received 21 May 2005; revised 8 December 2005; accepted 13 December 2005)

In psychoacoustic studies there is often a need to assess performance indices quickly and reliably. The aim of this study was to establish a quick and reliable procedure for evaluating thresholds in backward masking and frequency discrimination tasks. Based on simulations, four procedures likely to produce the best results were selected, and data collected from 20 naive adult listeners on each. Each procedure used one of two adaptive methods (staircase or maximum-likelihood estimation, each targeting the 79% correct point on the psychometric function) and two response paradigms (3-interval, 2-alternative forced-choice AXB or 3-interval; 3-alternative forced-choice oddball). All procedures yielded statistically equivalent threshold estimates in both backward masking and frequency discrimination, with a trend to lower thresholds for oddball procedures in frequency discrimination. Oddball procedures were both more efficient and more reliable (test–retest) in backward masking, but all four procedures were equally efficient and reliable in frequency discrimination. Fitted psychometric functions yielded similar thresholds to averaging over reversals in staircase procedures. Learning was observed across threshold-assessment blocks and experimental sessions. In four additional groups, each of ten listeners, trained on the different procedures, no differences in performance improvement or rate of learning were observed, suggesting that learning is independent of procedure. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2164988]

PACS number(s): 43.66.Yw, 43.66.Fe, 43.66.Dc [RAL] Pages: 1616–1625

I. INTRODUCTION

In studies of auditory perceptual learning (i.e., improvement in task performance resulting from practice), pre- and post-training performance is usually compared to quantify training-induced learning on both trained and transfer conditions (e.g., Delhommeau *et al.*, 2002; Demany, 1985; Demany and Semal, 2002; Wright *et al.*, 1997a). The pretests used often include hundreds of trials (e.g., Delhommeau *et al.*, 2002 used 900 pretraining trials; Wright *et al.*, 1997a used 1500–1800 trials). We have previously shown that a major portion of perceptual learning, at least in a frequency discrimination task, occurs early in training, within the first 500–1000 trials (Hawkey *et al.*, 2004). Thus, excessive pretesting prior to the start of a designated "training" session may actually provide a great deal of training, and interesting learning phenomena might be missed. In order to study these rapid learning effects, a method for assessing thresholds as quickly, accurately, and reliably as possible in naive listeners is needed.

Psychophysical procedures will be discussed following the taxonomy introduced by Marvit *et al.* (2003): the *procedure* is a combination of *paradigm* (the *mode of stimulus presentation* which defines the number of presented intervals and their content, and the *task* which defined what the listener has to do) and *method* (incorporating the *measurement strategy*, or rules used to select stimuli in the course of the measurement, and the *datum definition*, which defines the result of the measurement given the presented stimuli and the listener's responses to them). This study compared four procedures, produced by a combination of two paradigms and two methods, and addressed the issue of finding the appropriate datum definition.

Monte Carlo simulations were used to narrow down the range of possible procedures likely to produce best results. The simulations compared multi-interval maximum-likelihood (ML) estimation methods targeting either the 79% correct point, or the "sweetpoint" (point minimizing the variability in threshold estimates; see Green, 1990), on the psychometric function and a 3-down, 1-up staircase which also targeted the 79% correct point (Levitt, 1971). Staircases were implemented with or without a "lead-in" phase in which large steps were used with a 1-down, 1-up rule until the first incorrect response, at which point the 3-down, 1-up phase began (as suggested by Baker and Rosen, 2001). Datum definitions were trial signal value for the ML methods and averages over 2 or 4 last reversals for the staircase methods. For each method, 2- and 3-alternative, forced-choice (AFC) paradigms were compared. Six observers performing a discrimination task were simulated, with thresholds logarithmically spread over the range observed in human listeners for a frequency discrimination task (Amitay *et al.*, 2005). For each observer, psychometric functions defining the probability of a correct response as a function of the signal level were defined for the different paradigms based on signal-detection

[a)]Portions of this work were presented in "Rapid and reliable assessment of psychoacoustic thresholds," ARO Midwinter Meeting, New Orleans, Louisiana, 2005 [Abstract #1028].

[b)]Current affiliation: Developmental Cognitive Neuroscience Unit, Institute of Child Health-UCL, 30 Guilford Street, London WC1N 1EH, United Kingdom. Electronic mail: s.amitay@ich.ucl.ac.uk

 0001-4966/2006/119(3)/1616/10/$22.50

theory. In each simulated trial, a random number between 0 and 1 was generated and compared to the probability of a correct response as defined by the psychometric function, thus generating a correct (when < probability) or incorrect (when > probability) response. Signal level for the next trial was then selected using the measurement strategy for the procedure under investigation. Separate simulations were run for perfectly attentive observers and for inattentive ones. For inattentive observers, 3% of trials (randomly assigned) were "attention lapses," where the response was correct with chance probability.

The simulation results suggested that the ML method targeting the sweetpoint (94.2% for the 2AFC paradigm and 91.6% for the 3AFC paradigm, taking into account the logarithmic spread of stimulus values) coupled with a 3AFC paradigm was quickest to converge on the threshold and most accurate for perfectly attentive observers. However, this procedure failed when attentional lapses were introduced into the model. Under conditions of inattention, a 2AFC ML targeting a lower point on the psychometric function (79%) was more accurate and quicker to converge on the threshold. The staircase method coupled with a 2AFC paradigm appeared to be more robust to simulated variations in attention when a lead-in phase was used.

Since human listeners are often less than perfectly attentive, we decided to compare empirical data only for those methods promising best performance under these conditions: staircase and ML targeting the 79% correct point on the psychometric function, coupled with either a 2- or a 3AFC paradigm. In the 3-interval, 2AFC AXB listeners match the first or last interval (A or B) to the middle interval (X), and in 3-interval, 3AFC oddball listeners pick the "odd-one-out" (the interval that differs from the other two). Both paradigms require same–different comparisons, rather than an identification of signal level which requires verbal labeling of the stimuli (e.g., high–low in a 2AFC frequency discrimination task). Such labeling can be problematic for populations with language immaturity or impairments. Using the 3-interval design, the 2- and 3-AFC paradigms are equal for administration time, allowing a direct comparison of efficiency. In fact, both paradigms present two identical intervals and one that is different: in the 2AFC AXB the different one can be either the first or the last, whereas in the oddball paradigm it can be any one of the three intervals.

The methods compared in this study were ML (Green, 1993) and the more commonly used 3-down, 1-up staircase (Levitt, 1971) targeting the 79% correct point on the psychometric function, allowing for a direct comparison of the two. In the staircase method, knowledge of the exact shape of the psychometric function is unnecessary, but prior knowledge of its slope and range is useful in defining task-appropriate starting rules and step sizes. However, staircases are generally considered to require a large number of trials to achieve a robust threshold estimation (Baker and Rosen, 1998; Shelton *et al.*, 1982; but see Buss *et al.*, 2001). Moreover, resolution depends on step size, which cannot be made too small at the risk of very long blocks. By contrast, it is claimed ML requires as few as 12–15 trials to produce an accurate threshold (Florentine *et al.*, 2001; Green, 1993; though others find that typically twice as many trials are needed: Leek *et al.*, 2000), and resolution (i.e., the minimum difference between possible estimates) is as good as computation power allows without sacrificing length of testing block. The major drawbacks of the ML method are the need for prior knowledge of the shape of the psychometric function, the paucity of suprathreshold trials, and its susceptibility to attentional lapses, especially those that occur early in the block.

Several studies have compared multi-interval ML and staircase procedures (Baker and Rosen, 2001; Buss *et al.*, 2001; Marvit *et al.*, 2003; Saberi and Green, 1997; Shelton, *et al.*, 1982; Shelton and Scarrow, 1984). These studies varied in their implementation of staircase and ML and in other ways (below). However, most appear to agree that threshold estimates provided by the different procedures tend to be very similar. Differences between procedures were found in their reliability and efficiency, but the studies disagreed on the direction of these differences. Shelton *et al.* (1982) and Buss *et al.* (2001) claim the efficiency and within-subject variability of ML procedures are similar to staircase, while Shelton and Scarrow (1984) and Saberi and Green (1997) show smaller variability in ML than staircase, and Marvit *et al.* (2003) showed lower efficiency in ML compared to staircase. This disparity is, perhaps, unsurprising given the variation between studies in task, method of presentation, use of feedback or lack thereof, length of testing blocks, choice of listener population (i.e., using experienced or naive listeners) and the amount of prior practice they received. These parameters often varied even within a single study (e.g., Marvit *et al.* 2003).

Thus, despite the importance of choosing an appropriate measurement procedure, there appears to be little consensus as to what is the fastest and most efficient method for assessing psychoacoustic thresholds, especially in naive listeners. A systematic investigation of commonly used procedures, using an experimental design that allows them to be directly compared on speed, efficiency, and reliability, is in order.

Bearing in mind that different procedures might be "best" for different tasks, we carried out a systematic evaluation of procedures for the detection of backward-masked tones and for frequency discrimination. Large individual differences are usually observed in both these tasks (e.g., Amitay *et al.*, 2005; Buss *et al.*, 1999), so any significant differences between procedures (after averaging over a highly variable population) are more likely to reflect consistent (or "real") differences between them. This sensitivity to individual differences makes these tasks good tools in distinguishing clinical populations from normal cohorts (e.g., Wright *et al.*, 1997b), and therefore good candidates for exploration.

A second issue, as mentioned above, is one of datum definition. The most widely used method for extracting thresholds from staircase procedures is to average stimulus presentation levels over an even number of reversals (points at which the direction of signal level change reverses). ML, on the other hand, selects as its stimulus the level corresponding to the targeted percent correct on the most likely function fitted to the data collected on all previous trials, and the threshold is simply the level of that stimulus on the last

trial in the block. Thus, ML can be stopped at any time to produce a threshold, while staircase requires a certain minimum number of reversals to occur (a variable number of trials), making it difficult to compare their efficiency on a trial-by-trial basis. This difficulty can be overcome by using a common method of extracting thresholds from the data obtained from adaptive procedures. One such method is fitting a psychometric function to the data (see Wichmann and Hill, 2001a). This is essentially similar to the ML method for adaptive tracking, but can be applied to data obtained from any procedure, and allows for trial-by-trial comparison of ML and staircase data. While this method has been used for estimating psychoacoustic thresholds (e.g., Lutfi *et al.*, 2003; Wightman *et al.*, 2003; Witton *et al.*, 2000), it has been contended that it produces biases in the slope estimates of the psychometric function for short adaptive tracks (Kaernbach, 2001; Leek *et al.*, 1992). However, to our knowledge no attempt has been made to examine whether fitting methods really constitute an improvement over the traditional threshold estimation methods for empirical data.

A third issue, which is relevant to studies of learning, is the selection of the best procedure for optimizing training. In threshold estimation, the assumption is that the listener's threshold remains constant. However, in studies of training it is necessary to track a changing threshold as accurately as possible. The rate of change may differ between tasks. In this study we explored frequency discrimination training. If the threshold changes too fast, a staircase may be too slow-changing to drive (and observe) fast learning. On the other hand, the paucity of suprathreshold trials and the relative volatility in the earlier part of the block in ML may cause confusion that might affect learning. Since ML takes into account all previous responses, it too may not respond fast enough to changes in threshold level, especially on long testing blocks. We addressed the question of the efficacy of these procedures for training in frequency discrimination in further testing.

To summarize, a reliable measure for threshold estimation is required for studies in which only a single threshold estimate can be obtained (e.g., perceptual learning experiments). Speed of convergence on threshold is also a requirement (to minimize the number of trials used in measuring pretraining thresholds). Thus, the aims of the first and second parts of this study were to define the "best" psychophysical procedure for acquisition of rapid and reliable thresholds in naive listeners on a detection and a discrimination task, and to find the most reliable method for obtaining threshold estimates from the adaptive data (datum definition). The best procedure is defined on the basis of speed in converging on a threshold, efficiency, and test–retest reliability. In the third part, the best procedure for training was explored. Here, the best procedure was defined as the one producing the greatest improvement in performance over the shortest time.

II. GENERAL METHODS

A. Listeners

All listeners recruited for this study lacked previous experience in psychoacoustic experiments. They were recruited through advertisements and notices posted on public university notice boards, and paid for their participation. Listeners were only excluded from participation if they failed to pass audiometric screening (20 dB HL or less bilaterally, at 0.5, 1, 2, and 4 kHz), administered in accordance with the BSA standard, method A (British Society of Audiology, 1981). Informed consent was obtained from all participants. This study was approved by the University of Nottingham, Department of Psychology Ethics Committee.

B. Apparatus

Stimuli were digitally generated using custom software running on a PC which also controlled the experiment. The signal waveforms were computed on each presentation using a sampling rate of 44.1 kHz, and were generated as 16-bit samples using a sound card (Darla Echo; Echo Digital Audio Corporation, Carpinteria, CA). Stimuli were presented diotically via Sennheiser HD 25-1 headphones. Listeners were tested individually in a sound-attenuating chamber and responded on a touch screen, which also visually cued the test intervals and provided visual feedback.

C. Stimuli

1. Backward masking

The test interval contained a 20-ms (10-ms $\cos^2$ ramp, no steady state), 1-kHz tone followed by a 300-ms (10 -ms $\cos^2$ ramp), 40-dB SPL (overall level) bandpass noise masker (center frequency: 1 kHz, width: 800 Hz) with no delay between tone and masker. The tone level was varied adaptively. The standard interval contained only the masker. The intervals were separated by 580 ms.

2. Frequency discrimination

All tones were 100 ms (10-ms $\cos^2$ ramp), separated by 500-ms interstimulus intervals, presented at 70-dB SPL. The standard was a 1-kHz tone, and the test tone was always higher, with the frequency difference ($\Delta F/F$, expressed in percent) between standard and test tone varied adaptively.

D. Procedures

The four procedures were the four possible combinations of adaptive method and paradigm (described below): staircase oddball, staircase AXB, ML oddball, and ML AXB. Testing blocks were terminated after 40 trials.

1. Adaptive methods

a. Staircase. In the backward-masking task, the initial tone level was 90 dB SPL. Different step sizes were used over the course of each run to speed up approach on the region of the threshold. The initial step size was 10 dB, implemented in a 1-up, 1-down manner for three reversals, then 5 dB for one more reversal. In the 3-down, 1-up phase the step size was 3 dB. During this phase, two consecutive steps in the same direction caused the step size to be multiplied by a factor of 1.5, reverting back to 3 dB when a reversal occurred. In the frequency discrimination task, the initial ΔF was 50% (500 Hz). During the 1-down, 1-up phase,

ΔF was halved after each step, until the first error occurred. The staircase then moved into a 3-down, 1-up phase where ΔF was multiplied or divided by a factor of $\sqrt{2}$. During this phase, two consecutive steps in the same direction caused ΔF to be multiplied (or divided) by a factor of 2, reverting back to $\sqrt{2}$ when a reversal occurred.

b. Maximum-likelihood (ML). The signal level presented on each trial was the estimated 79% correct point of the psychometric function which best fitted the data collected on the block up to that point (see Green, 1995 for details). Psychometric curves were generated based on signal-detection theory for the appropriate number of choices used (3 for the oddball and 2 for the AXB). For each task, 180 functions were generated to cover the range from chance to asymptotic performance, encompassing the range of thresholds measured in children and adults found in the literature (backward masking: Buss *et al.*, 1999; Hartley *et al.*, 2000; Roth *et al.*, 2001; frequency discrimination: Thompson *et al.*, 1999; Ahissar *et al.*, 2000; Amitay *et al.*, 2005). For backward masking the functions were equally (linearly) spaced, covering the range 0–90 dB SPL, and had a fixed slope of 0.4 (see Buss *et al.*, 2001), and for frequency discrimination the functions were logarithmically spaced, covering the range 0.1%–50% frequency difference from standard, and had a slope of 1 (see Saberi and Green, 1997). The slopes were fixed to simplify computation, since it has been shown that threshold estimates are relatively insensitive to small variations in the slope parameter (e.g., Green, 1990). Asymptotic performance was set at 100%, as the computer simulations showed that unless listeners were highly inattentive this will have little effect on threshold estimates.

2. Response paradigms

a. Oddball. Three intervals were presented: one containing the test stimulus and two containing the standard. Listeners decided which interval contained the test stimulus (3I-3AFC).

b. AXB. Three intervals were presented. Either the first (A) or the last (B) interval was identical to the middle interval (X), which contained either the standard or the test stimulus. Listeners decided whether the first or the last stimulus matched the middle one (3I-2AFC).

III. THRESHOLD ESTIMATION IN NAÏVE LISTENERS

A. Methods

1. Listeners

Twenty adult listeners with no previous experience in psychoacoustic testing were recruited for this experiment (10 male, 10 female; aged 19–30 years, mean age 22.2 years).

2. Experimental procedure

A testing session generally took 1.5–2 h, and included an audiometric screening test and 16 blocks of threshold assessments, 8 in each task (2 each per procedure). Listeners could take a break at any time, but were also required to take a break halfway through a session (after completing the first eight blocks). Before each switch in task and response paradigm, a short demonstration (five trials) was presented to acquaint the listener with the task and response demands. The demo included three trials where the test and standard stimuli were easily discriminable and two trials where the stimuli were not easily discriminable. Listeners were instructed to guess when they could not tell the difference. Procedure and task order were pseudorandomized across listeners, to avoid consistent differences between procedures due to learning.

3. Data analysis

For each trial, a logistic psychometric function [percentage of correct responses as a function of the adaptive parameter value; see Eq. (1)] was fitted to the pooled data collected up to and including that trial, using the fitting technique described by Wichmann and Hill (2001a; 2001b). Different types of psychometric function shapes (e.g., Weibull functions) were compared during initial analyses, and logistic functions yielded the best overall fits to our data. The probability of responding correctly when the adaptive parameter value is x is

$$\Psi(x) = \gamma + \frac{1 - \gamma - \lambda}{1 + e^{-(x/\alpha)^{\beta}}} \tag{1}$$

where α is the midpoint of the psychometric function (the parameter value corresponding to the halfway point between chance and asymptotic performance), β is the slope parameter, γ is the guessing rate (chance performance level), and λ is the asymptotic performance level, used to estimate the lapse rate. α was constrained to be within "reasonable" psychophysical limits (−10–100 dB in backward masking, 0.01%–50% in frequency discrimination). β and λ were both unconstrained, and γ was fixed at 0.5 for the 2AFC paradigm and at 0.33 for the 3AFC paradigm.

Monte Carlo simulations were used to obtain the confidence limits for the goodness-of-fit measure. The psychometric function fitted to the empirical data was used as a generating function, and 1999 data sets were simulated (Wichmann and Hill, 2001a). The goodness of fit (root-mean square deviance between data and fitted function) was assessed for each simulated set. The distribution of the goodness-of-fit measure yielded the confidence limits. The fit was considered a failure if the estimated goodness of fit obtained from the empirical data fell outside the 90% confidence limits (i.e., if it was larger than the simulated goodness-of-fit measures from 95% of the data sets). Attempts to fit a function to less than 20 data points often failed to converge, and in many cases failed the goodness-of-fit test; therefore, the 79% correct threshold (hereafter "threshold") for each trial was extracted from the 20th onwards. The rate of fit failures was used as a criterion to compare datum definitions.

Thresholds obtained from fitting psychometric functions to the data were compared to those obtained in the "traditional" way: the adaptive parameter value on the last trial in the ML, and averaging over the last two or four reversals in the staircase method. Four, and occasionally even two, reversals did not always occur, as the stop criterion was a fixed

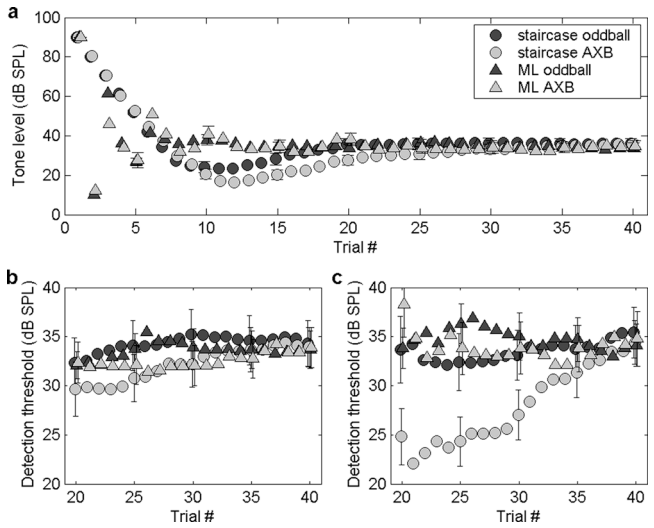


FIG. 1. Backward-masking detection thresholds, as a function of the trial at which the measurement was taken: (a) Group averages of the actual tone levels presented during the experiment; (b) 79% correct thresholds obtained from fitting logistic psychometric functions to the data, averaged across blocks and listeners. The threshold plotted for each trial is based on all data points collected up to and including that trial. (c) Thresholds obtained from traditional analyses: for ML procedures, this is the adaptive parameter value presented on that trial; for staircase procedures, it is the arithmetic mean of most recent two reversals that occurred prior to and including that trial. Error bars are standard errors of the means. The errors remained unchanged throughout and for clarity are thus plotted for selected trials only.

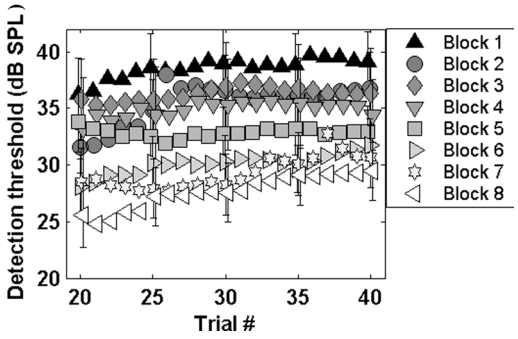


FIG. 2. Mean backward-masking thresholds (from psychometric function fits) averaged across procedures and plotted as a function of trial number (psychometric functions fitted to data accumulated up to and including that trial). Block are plotted in order of presentation. Error bars are standard errors of the means. The errors remained unchanged throughout and for clarity are thus plotted for selected trials only.

number of trials. Since previous work has shown that there is little loss in accuracy in averaging over two rather than four reversals (or more; Baker and Rosen, 2001), data analysis was limited to two reversals. Thresholds were calculated only over those blocks where reversals had occurred, and the proportion of blocks that failed to yield a threshold was used as a further criterion to compare datum definitions. In the case of the ML, the parameter value was not identical to the threshold estimate from the fitted function, as the functions were defined differently, and the slope was not constrained in the fitted measure as it was in the adaptive procedure. All statistical analysis was carried out on the fitted data and not the traditional method because too many data points were missing (two reversals failed to occur by trial 20 on many blocks).

The frequency discrimination data violated the normality assumption because thresholds were skewed toward low frequencies, and therefore were log-transformed for the purpose of statistical analysis.

B. Results

1. Backward masking

a. Threshold analysis. Threshold estimates obtained using psychometric function fits and traditional analysis methods for the backward masking data are shown in Figs. 1(b) and 1(c), respectively. Using both forms of analysis, all procedures asymptote to the same threshold (within experimental error). The staircase AXB (gray circles) tended to go well below threshold at the beginning of the testing block [Fig. 1(a)], resulting in a distinct trend for the estimated thresholds based on reversals to rise over time [Fig. 1(c)]. This problem was somewhat less marked when fitting the data, where later trials (signal levels closer to threshold levels) influenced the placement of the fitted psychometric function even in the absence of further reversals. A repeated-measures analysis of variance (ANOVA) of thresholds from the fitted data [Fig. 1(b)], with order of presentation as a covariate, revealed a significant main effect of trial [$F(20,3100)=1.85$, $p=0.012$] as thresholds drifted upwards by approximately 2–4 dB from trial 20 to trial 40. This drift was more pronounced for the staircase procedures than the ML [trial × paradigm interaction: $F(20,3100)=2.19$, $p=0.002$]. However, there were no significant main effects of either adaptive method [$F(1,155)=0.17$, $p=0.685$] or response paradigm [$F(1,155)=0.23$, $p=0.634$]. The only other significant effect was test block presentation order [$F(1,155)=11.42$, $p=0.001$]. This effect is illustrated in Fig. 2, and suggests that learning in the backward-masking task occurred extremely rapidly; thresholds decreased by approximately 10 dB from the first to the eighth block, or 320 trials.

An important issue in fitting psychometric functions is the goodness of fit to the data. Two percent of all data points (single trials) failed a test for goodness of fit (see the full description in Sec. III A 3 above). This proportion of "failed" fits is within the acceptable statistical error limit, so it was not taken into account in the above analysis. Over 50% of failed fits occurred prior to trial 25, and almost 80% of them prior to trial 30. This suggests if 30 trials are used in acquiring the data it can be reliably fitted, with only about 0.4% probability of a failed fit.

The traditional analysis yielded similar thresholds to the psychometric function fits, except for the staircase AXB procedure. However, the threshold estimate at each trial for the two staircase procedures in Fig. 1(c) includes only on those blocks where two reversals had occurred by that trial (e.g., data at trial 30 is the mean over threshold estimates in all blocks where the second reversal occurred on or prior to trial 30). In over 50% of all blocks, two reversals did not occur by trial 20. Therefore, full statistical analysis was not possible for those data. The strong downward deflection seen in the reversal averaging in the staircase AXB procedure cannot be

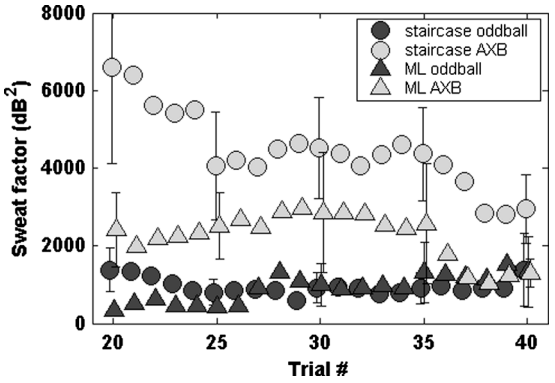


FIG. 3. Sweat factors (see the text) associated with the different procedures for the backward-masking task. Error bars are standard errors of the means.

a result of insufficient data points, as the first two reversals occurred at similar points for the staircase AXB and staircase oddball procedures.

ML oddball yielded the smallest difference (0.1–0.2 dB on average) between estimated thresholds on trial 25 and trial 40 [which was taken to be the asymptotic threshold estimate; Fig. 1(b)]. This difference in threshold estimates did not decrease substantially when increasing the number of trials from 25 to 30.

b. Sweat factors. The sweat factor is defined as the product of the within-subject variance in threshold estimation and the number of trials (see Marvit *et al.*, 2003; Taylor and Creelman, 1967; Taylor *et al.*, 1983). The higher the sweat factor, the less efficient the procedure. The sweat factors associated with the tested procedures on the backward-masking task are shown in Fig. 3. The oddball procedures had significantly lower sweat factors than the AXB procedures [$F(1,76)=7.92$, $p=0.006$], but there was no significant difference between staircase and ML.

2. Frequency discrimination

a. Threshold analysis. Four blocks (out of a total of 320) were excluded from analysis: three from the ML AXB procedure, where errors on the first trial prevented the block from converging on a threshold, and one from the staircase oddball procedure, where an error in the early part of the block had the same effect (this problem can be easily avoided in the future by stopping the test block if there is an error in the first trial). Figure 4(b) and 4(c) show threshold estimates for frequency discrimination for psychometric function fits and traditional analysis, respectively. The difference between analysis methods appears marginal, except for the later trials in the staircase oddball procedure, where there is a trend for the thresholds to drift up when averaged over the last two reversals. Since fitting the data takes into account all previous trials, rather than the two last reversals only, it is less susceptible to local variations in performance. The thresholds obtained from the oddball paradigm were slightly lower than those from the AXB paradigm, but not significantly so [$F(1,151)=1.26$, $p=0.263$]. The order effect was not significant either [$F(1,151)=2.78$, $p=0.097$], though thresholds on the first track presented were generally higher than following tracks (see Fig. 5). This suggests that the time

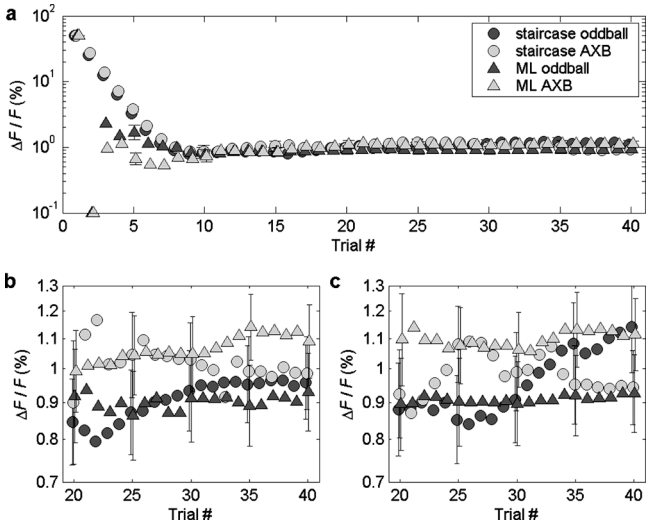


FIG. 4. Frequency discrimination thresholds plotted in the same format as Fig. 1: (a) Group averages of the frequency difference; (b) thresholds from psychometric function fits; and (c) thresholds from traditional analyses. Error bars are standard errors of the means, plotted in the same manner as Fig. 1.

course of learning is different in frequency discrimination and backward masking (see Watson, 1980). No other effects were significant. The psychometric function fits failed the goodness-of-fit test at 3.2% of all data points. Again, 50% of these bad fits occurred prior to trial 25, and almost 80% before trial 30, suggesting at least 30 trials are necessary for robust data fitting. These bad fits usually result either from errors on suprathreshold stimuli, or from a run of correct responses on subthreshold stimuli, specially if those occur early in the block. In over 98% (all but 1) of the blocks the second reversal had occurred by trial 30. Differences in threshold estimates between trial 25 and 40 were not large (generally less than 1 Hz), and were not substantially reduced by increasing the number of trials from 25 to 30.

b. Sweat factors. No significant differences in the sweat factor were observed (Fig. 6) between procedures beyond trial 30 (ANOVA yielded no significant effects).

C. Discussion

Neither backward masking nor frequency discrimination threshold estimates differed significantly across the different procedures. However, oddball procedures (ML and staircase)

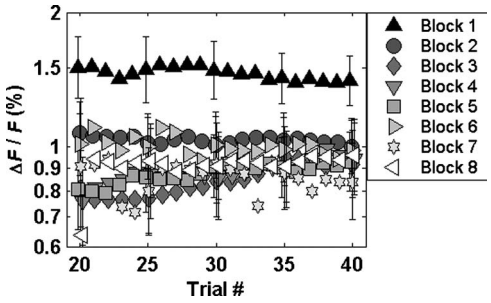


FIG. 5. Frequency discrimination thresholds, averaged across procedures, plotted in a similar format to Fig. 2. Error bars are standard errors of the means.

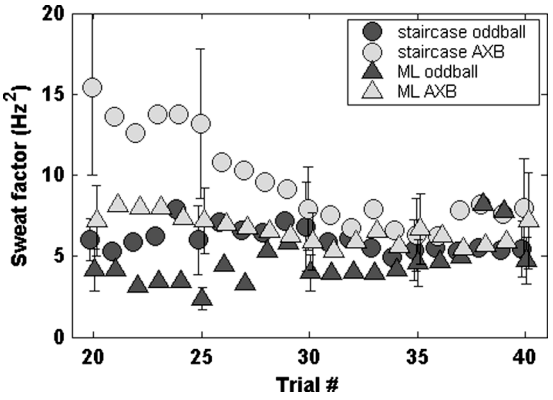


FIG. 6. Sweat factors for frequency discrimination, plotted in the same format as Fig. 3. Error bars are standard errors of the means. Note that while frequency discrimination thresholds were measured in percent of the standard tone frequency (1 kHz), these have been converted to actual frequency difference (in Hz) for the purpose of calculating the sweat factor.

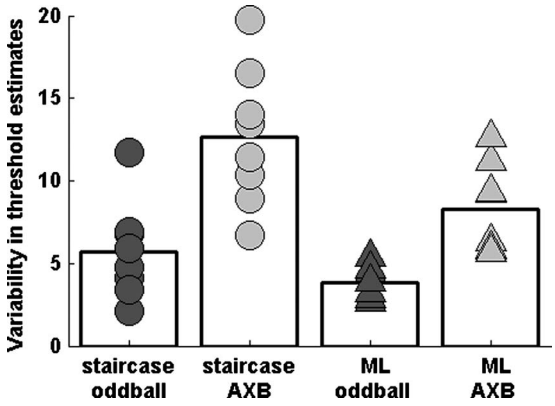


FIG. 7. Variability in threshold estimates (standard deviation of the six estimates obtained for each listener) for backward masking on each of the four procedures, averaged across listeners (bars). Individual data points are plotted as filled symbols.

provided a more efficient way than AXB of assessing thresholds in the backward-masking task. The efficiency of the backward-masking staircase oddball procedures may be improved further by removing additional reversals at large step sizes during the 1-up, 1-down phase. In the frequency discrimination, no differences in efficiency (i.e., sweat factors) were apparent beyond trial 30. Approximately 30 trials were necessary (and sufficient) to obtain accurate and efficient threshold estimates in naive listeners in both tasks. If the reliability of the fitting technique is deemed sufficient with 1%–1.5% failures, then 25 trials are sufficient to obtain a threshold using either the staircase oddball or the ML oddball procedure. This is similar to the 24 trials advocated by Leek *et al.* (2000) for a yes–no ML procedure. However, for more reliable threshold estimates, 30 trials are suggested.

For the ML oddball paradigm, fitting psychometric functions to the data did not differ significantly from using the adaptive parameter value, suggesting that fixing the slope of the psychometric function when acquiring the data did not have an adverse effect on threshold estimation, and that either method can be used for datum definition if a ML procedure is used. For the staircase method, fitting a psychometric function to the data had the advantage that a threshold could be estimated after 30 trials, whereas more than 30 trials were occasionally necessary to get sufficient reversals for averaging (especially in the backward-masking task). Thus, threshold estimates based on fitting psychometric functions to the data were the better datum definition for staircase procedures, and showed no disadvantage compared to the traditional estimates for ML procedures. There is also indication that the exact shape of the function is relatively unimportant, as there is little difference in estimated threshold between using a cumulative Gaussian distribution (based on signal detection theory) in the adaptive method and a logistic function in fitting the data. Thus, any biases that may have existed in the slope of the psychometric function (Kaernbach, 2001), when fitted to data from the adaptive procedures, had no significant effect on threshold estimation. Therefore, we suggest that threshold estimates can be validly obtained from fitting the data when psychometric slopes are not a necessary part of the datum definition.

IV. TEST–RETEST RELIABILITY

A. Methods

1. Listeners

Ten of the listeners who took part in the first experiment (5 male, 5 female; aged 20–30 years, mean age 24.0 years) were asked to return 1 week (±1 day) and 3 weeks (±2 days) following initial testing to assess test–retest reliability. These listeners were self-selected; all participants in the first experiment were given the choice of returning for further testing. One of the ten listeners failed to attend the third session. A second listener produced multiple threshold assessment failures during the third session and was therefore excluded from this experiment.

2. Experimental procedure

Both additional sessions were identical to the first experiment, except that the audiometric screening test was not repeated.

3. Data analysis

The first experiment showed that 30 trials were sufficient for a stable threshold estimate. Thresholds were therefore calculated in this experiment using psychometric function fits to the data accumulated up to and including trial 30. The within-subject standard deviation was calculated over the six threshold estimates for each procedure and averaged over listeners for a measure of the mean variability. The higher the variability, the lower the test–retest reliability associated with the procedure.

B. Results

Altogether six threshold estimates were taken for each procedure and for each task. The within-subject variability associated with each procedure was assessed as the standard deviation over the six threshold estimates (two per session). Group mean variability for the procedures in the backward-masking task is shown in Fig. 7 (bars), along with the individual within-subject variability estimates (symbols). The variability associated with oddball procedures was consider-

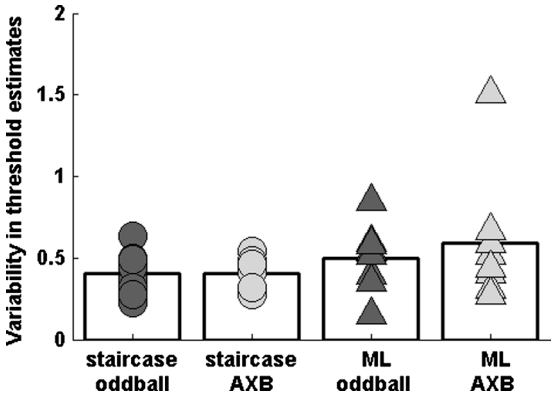


FIG. 8. Variability in threshold estimates (standard deviation of estimates) for frequency discrimination, plotted in the same format as Fig. 7.

ably lower than the AXB procedures. The variability in ML procedures was also slightly lower compared to the staircase. The within-subject test–retest standard deviation was examined with a two-way ANOVA, with method and paradigm as the factors. Oddball paradigms had lower variability than AXB [$F(1,28)=30.02$, $p<0.001$], and ML threshold estimates had lower variability than staircase [$F(1,28)=9.05$, $p=0.005$]. Learning was demonstrated by an overall trend for thresholds to decrease with repeated tests [$F(5,135)=6.45$, $p<0.001$; data not shown], but interaction terms with either method or paradigm were not significant, suggesting the decreases in thresholds followed similar time courses for the different procedures. The lack of difference in learning effects between procedures is not surprising, since each listener was tested on all procedures in each session. It was previously shown that perceptual learning transfers across procedures (Hawkey *et al.*, 2004), so even if a particular procedure was more effective for learning, it will contribute to performance improvement seen for all other procedures.

Group mean standard deviations for the frequency discrimination test-retest are shown in Fig. 8 (bars), along with the individual within-subject standard deviations (symbols). There were no significant differences between procedures. As for backward masking, significant learning was observed as a trend for decreasing thresholds [$F(5,140)=2.59$, $p=0.029$; data not shown].

C. Discussion

In the backward-masking task, both oddball procedures showed higher test–retest reliability (lower variability) than AXB procedures. In frequency discrimination, the reliability of the different procedures was equivalent. The changes observed in the thresholds over time (learning) increased the variability, but since no differences were found between procedures in the time course or magnitude of these changes they are unlikely to be the cause of the difference between procedures in the backward-masking task. The results of the two experiments suggest that the oddball paradigm, used with either a staircase or ML method, is a quick, efficient, and reliable procedure for estimating thresholds in both the backward-masking and frequency discrimination tasks.

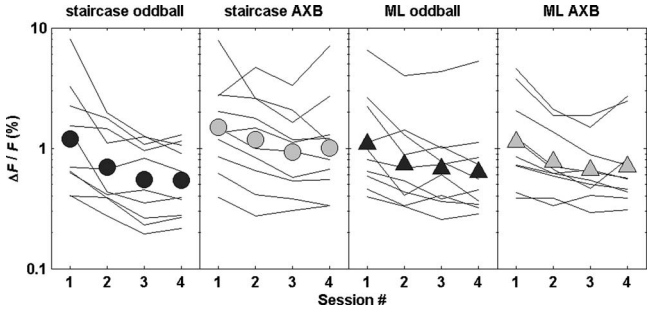


FIG. 9. Frequency discrimination thresholds plotted as a function of session. Each data point is the geometric mean of eight threshold estimates, obtained from psychometric functions fitted to data from all 40 trials. Individual data are shown as lines, group mean data as filled symbols.

V. COMPARISON OF PROCEDURES IN TRAINING

A. Methods

1. Listeners

Forty listeners with no prior experience with psychoacoustic testing (15 male, 25 female; aged 18–38 years, mean age 23.3 years) were recruited for this experiment. They were divided into four groups, ten listeners in each group. These listeners were recruited in the same way as the previous experiment, and the same exclusionary criteria applied.

2. Experimental procedure

Each group trained on frequency discrimination using one of the four procedures: group 1 trained on staircase oddball, group 2 on staircase AXB, group 3 on ML oddball, and group 4 on ML AXB. Training was conducted in four sessions, with no more than 3 days between sessions. Each testing session took approximately half an hour. The first session included an audiometric screening test and eight blocks of threshold assessments, each of 40 trials. Listeners could take a break at any time. At the beginning of each session, a short demonstration was presented to acquaint the listener with the task and response demands or to remind him/her of them. As before, the demo included three trials where the test and standard stimuli were easily discriminable and two trials where the stimuli were not easily discriminable. The procedures were implemented in the same way as in the first experiment.

3. Data analysis

The data from each block (all 40 trials) were fitted with a psychometric function. In each session, the eight threshold estimates were geometrically averaged to produce a single session-threshold estimate.

B. Results and discussion

Threshold estimates for the four groups over the four testing sessions are shown in Fig. 9. Individual data are shown as lines, and group averages as filled symbols. While the thresholds themselves differed between groups (or procedures—there is no way to differentiate the two), the rate of change of threshold and overall improvement appeared to be similar across groups. A repeated measures

ANOVA revealed significant learning [main effect of session; $F(3,108)=39.9$, $p<0.001$]. The effect of procedure (group) was not significant, and neither was the interaction between session and procedure, suggesting learning did not differ between procedures. Based on these results all procedures tested have the same efficacy for training and there is no learning-related reason to choose ML over staircase procedures, or vice versa.

VI. GENERAL DISCUSSION

This study was designed to compare common procedures for obtaining backward-masking and frequency discrimination thresholds in naive listeners, and to choose the one that is optimal in terms of speed, efficiency, and test–retest reliability. The effectiveness of the same procedures for training in frequency discrimination was also compared. We compared two methods—staircase and ML, implemented in two paradigms—3I-3AFC oddball and 3I-2AFC AXB.

In the backward-masking task, oddball procedures were consistently superior to AXB procedures on every measure, though in the frequency discrimination task no significant differences were found. This may be due to the high between- as well as within-subject variability usually observed in frequency discrimination, even when compared to backward masking. The 3AFC advantage over the 2AFC procedures supports the conclusion of Kollmeier *et al.* (1988), at least for the staircase procedures (ML procedures were not tested in that study), and extends it from experienced to naive listeners. However, while Shelton and Scarrow (1984) found ML to yield more consistent results in naive listeners, Marvit *et al.* (2003) concluded ML was less efficient and, like the current study, Buss *et al.* (1999) found no difference in efficiency between ML and staircase in children and adults. Bearing in mind that these studies compared procedures on different psychoacoustic tasks, further differences between these studies are in the details: Shelton and Scarrow (1984) tested the different procedures on different groups of listeners, so could not directly compare within-subject variability between procedures. Marvit *et al.* (2003), on the other hand, may not have used the most efficient implementation of ML, in either number of test trials (15 or 50). As was shown here, 15 trials are probably too few to produce stable and reliable threshold estimates in naive listeners, resulting in high variance, while 50 may be well beyond the point at which the variance is no longer reduced for ML. The small variability found for the staircase procedure was the result of 85 trials on average, used in an interleaved manner so that each threshold estimate was an average over three independent estimates, with outliers removed from the data. Their conclusion that staircase is the most efficient method holds only insofar as such prolonged testing blocks can be implemented. A different conclusion may have been reached if the methods were compared over an equal number of trials, as was done in the study reported here.

When more than 30 trials were used in threshold assessment, there was no difference in either threshold estimate, efficiency, or reliability between the ML- and staircase oddball procedures. The estimated threshold from 30 data points was as accurate (within experimental error) as that from longer blocks. Thus, for naive listeners, 30 trials are recommended for threshold assessment in both frequency discrimination and backward masking. If testing constraints (e.g., assessing pretraining thresholds, using large test batteries) dictate short blocks, ML can provide estimates that are nearly as accurate and only slightly less reliable (in test–retest) for 25-trial blocks as for longer testing blocks. For 25-trial blocks, however, there is the risk of an occasional (2%–3% of blocks) failure to estimate a threshold with the accumulated data. Using fewer than 25 trials is not recommended, as data are not sufficient to ensure a reliable threshold estimate. The testing-block length recommended here is significantly shorter than the 85-trial (on average) staircase advocated by Marvit *et al.* (2003), and even the 40–50 trials recommended by Shelton and Scarrow (1984). If using ML, the datum definition can be either the adaptive parameter value presented on the last trial, or threshold estimated from fitting a psychometric function to the data (where no constraint on the slope was applied). Staircase oddball can also be used in shorter blocks, but in that case datum definition should be based on fitting psychometric functions to the data, as a minimal number of reversals for averaging is not guaranteed.

The conclusion that 30 trials are necessary for a reliable threshold estimate holds so long as the region of the threshold is unknown, in this case when testing naive listeners who show remarkable interindividual variability. If the range of possible threshold values is known *a priori* (e.g., pilot data are available), fewer trials can be used to estimate thresholds reliably.

The results of this study show that many of the concerns relating to either ML or staircase mentioned in the Introduction are unfounded when using an oddball paradigm. The limited resolution of the staircase paradigm, which depended on the small step size, and the higher resolution of the ML yielded similar threshold estimates when the data were fitted with psychometric functions. Nor did the staircase method require longer blocks to achieve a robust threshold estimate—efficiency was about equal for blocks 30 trials or longer. The small number of suprathreshold trials in the ML did not seem to have an adverse effect on threshold estimation in naive listeners. On the other hand, the ML method was indeed susceptible to attentional lapses, especially at the beginning of a block; in this case the algorithm failed to converge on a threshold (see also Baker and Rosen, 2001). This problem can, however, be easily overcome by stopping the assessment if such an error occurs and restarting the testing block. Another option is creating a "hybrid" procedure by adding a short 1-up, 1-down staircase until the first error occurs, then starting the normal ML. Preliminary testing with this hybrid procedure has produced very promising results, and appears to yield extremely stable and reliable results in just 25 trials.

All tested procedures were equally effective for training, following similar time courses and resulting in comparable performance improvement. This is further evidence that perceptual learning is largely independent of procedure (see Hawkey *et al.*, 2004). The learning effects were observed

across rather than within blocks, suggesting that even such short, 40-trial blocks provided significant training on the task and that the threshold estimates provide a reliable measure of momentary performance.

ACKNOWLEDGMENTS

The authors would like to thank Silvia Cirstea for advice on statistical analysis, and Tim Folkard for the software package for psychophysical testing. Many thanks to Michael Akeroyd and Stuart Rosen for helpful comments on an earlier version of this manuscript. This research was entirely supported by the Medical Research Council, United Kingdom.

Ahissar, M., Protopapas, A., Reid, M., and Merzenich, M. M. (**2000**). "Auditory processing parallels reading abilities in adults," Proc. Natl. Acad. Sci. U.S.A. **97**, 6832–6837.

Amitay, S., Hawkey, D. J. C., and Moore, D. R. (**2005**). "Auditory frequency discrimination learning is affected by stimulus variability," Percept. Psychophys. **67**, 691–698.

Baker, R. J., and Rosen, S. (**1998**). "Minimizing boredom by maximizing likelihood—Efficient estimation of masked thresholds," Br. J. Audiol. **32**, 104–105.

Baker, R. J., and Rosen, S. (**2001**). "Evaluation of maximum-likelihood threshold estimation with tone-in-noise masking," Br. J. Audiol. **35**, 43–52.

British Society of Audiology (**1981**). "Recommended procedures for pure-tone audiometry using a manually operated instrument," Br. J. Audiol. **15**, 213–216.

Buss, E., Hall, J. W. III, Grose, J. H., and Dev, M. B. (**1999**). "Development of adult-like performance in backward, simultaneous, and forward masking," J. Speech Lang. Hear. Res. **42**, 844–849.

Buss, E., Hall, J. W., Grose, J. H., and Dev, M. B. (**2001**). "A comparison of threshold estimation methods in children 6–11 years of age," J. Acoust. Soc. Am. **109**, 727–731.

Delhommeau, K., Micheyl, C., Jouvent, R., and Collet, L. (**2002**). "Transfer of learning across durations and ears in auditory frequency discrimination," Percept. Psychophys. **64**, 426–436.

Demany, L. (**1985**). "Perceptual learning in frequency discrimination," J. Acoust. Soc. Am. **78**, 1118–1120.

Demany, L., and Semal, C. (**2002**). "Learning to perceive pitch differences," J. Acoust. Soc. Am. **111**, 1377–1388.

Florentine, M., Marvit, P., and Buus, S. (**2001**). "Maximum-likelihood yes–no procedure for gap detection: Effect of track length," J. Am. Acad. Audiol. **12**, 113–120.

Green, D. M. (**1990**). "Stimulus selection in adaptive psychophysical procedures," J. Acoust. Soc. Am. **87**, 2662–2674.

Green, D. M. (**1993**). "A maximum-likelihood method for estimating thresholds in a yes–no task," J. Acoust. Soc. Am. **93**, 2096–2105.

Green, D. M. (**1995**). "Maximum-likelihood procedures and the inattentive observer," J. Acoust. Soc. Am. **97**, 3749–3760.

Hartley, D. E., Wright, B. A., Hogan, S. C., and Moore, D. R. (**2000**). "Age-related improvements in auditory backward and simultaneous masking in 6- to 10-year-old children," J. Speech Lang. Hear. Res. **43**, 1402–1415.

Hawkey, D. J. C., Amitay, S., and Moore, D. R. (**2004**). "Early and rapid perceptual learning," Nat. Neurosci. **7**, 1055–1056.

Kaernbach, C. (**2001**). "Slope bias of psychometric functions derived from adaptive data," Percept. Psychophys. **63**, 1389–1398.

Kollmeier, B., Gilkey, R. H., and Sieben, U. K. (**1988**). "Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model," J. Acoust. Soc. Am. **83**, 1852–1862.

Leek, M. R., Hanna, T. E., and Marshall, L. (**1992**). "Estimation of psychometric functions from adaptive tracking procedures," Percept. Psychophys. **51**, 247–256.

Leek, M. R., Dubno, J. R., He, N., and Ahlstrom, J. B. (**2000**). "Experience with a yes–no single-interval maximum-likelihood procedure," J. Acoust. Soc. Am. **107**, 2674–2684.

Levitt, H. (**1971**). "Transformed up–down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Lutfi, R. A., Kistler, D. J., Callahan, M. R., and Wightman, F. L. (**2003**). "Psychometric functions for informational masking," J. Acoust. Soc. Am. **114**, 3273–3282.

Marvit, P., Florentine, M., and Buus, S. (**2003**). "A comparison of psychophysical procedures for level-discrimination thresholds," J. Acoust. Soc. Am. **113**, 3348–3361.

Roth, D. A., Kishon-Rabin, L., and Hildesheimer, M. (**2001**). "Auditory backward masking and the effect of training in normal hearing adults," J. Basic Clin. Physiol. Pharmacol. **12**, 145–159.

Saberi, K., and Green, D. M. (**1997**). "Evaluation of maximum-likelihood estimators in nonintensive auditory psychophysics," Percept. Psychophys. **59**, 867–876.

Shelton, B. R., and Scarrow, I. (**1984**). "Two-alternative versus three-alternative procedures for threshold estimation," Percept. Psychophys. **35**, 385–392.

Shelton, B. R., Picardi, M. C., and Green, D. M. (**1982**). "Comparison of three adaptive psychophysical procedures," J. Acoust. Soc. Am. **71**, 1527–1533.

Taylor, M. M., and Creelman, C. D. (**1967**). "PEST: Efficient estimates of probability functions," J. Acoust. Soc. Am. **41**, 782–787.

Taylor, M. M., Forbes, S. M., and Creelman, C. D. (**1983**). "PEST reduces bias in forced-choice psychophysics," J. Acoust. Soc. Am. **74**, 1367–1374.

Thompson, N. C., Cranford, J. L., and Hoyer, E. (**1999**). "Brief-tone frequency discrimination by children," J. Speech Lang. Hear. Res. **42**, 1061–1068.

Watson, C. S. (**1980**). "Time course of auditory perceptual learning," Ann. Otol. Rhinol. Laryngol. Suppl. **89**, 96–102.

Wichmann, F. A., and Hill, N. J. (**2001a**). "The psychometric function. I. Fitting, sampling, and goodness of fit," Percept. Psychophys. **63**, 1293–1313.

Wichmann, F. A., and Hill, N. J. (**2001b**). "The psychometric function. II. Bootstrap-based confidence intervals and sampling," Percept. Psychophys. **63**, 1314–1329.

Wightman, F. L., Callahan, M. R., Lutfi, R. A., Kistler, D. J., and Oh, E. (**2003**). "Children's detection of pure-tone signals: Informational masking with contralateral maskers," J. Acoust. Soc. Am. **113**, 3297–3305.

Witton, C., Green, G. G. R., Rees, A., and Henning, G. B. (**2000**). "Monaural and binaural detection of sinusoidal phase modulation of a 500-Hz tone," J. Acoust. Soc. Am. **108**, 1826–1833.

Wright, B. A., Buonomano, D. V., Mahncke, H. W., and Merzenich, M. M. (**1997a**). "Learning and generalization of auditory temporal-interval discrimination in humans," J. Neurosci. **17**, 3956–3963.

Wright, B. A., Lombardino, L. J., King, W. M., Puranik, C. S., Leonard, C. M., and Merzenich, M. M. (**1997b**). "Deficits in auditory temporal and spectral resolution in language-impaired children," Nature **387**, 176–178.

Effects of auditory feedback on fricatives produced by cochlear-implanted adults and children: Acoustic and perceptual evidence

Sneha V. Bharadwaj, Emily A. Tobey, Peter F. Assmann, and William F. Katz
School of Behavioral and Brain Sciences, University of Texas at Dallas,
Callier Center for Communication Disorders, 811 Synergy Park Blvd. Richardson, Texas 75080

(Received 17 December 2004; revised 19 December 2005; accepted 28 December 2005)

Acoustic analyses and perception experiments were conducted to determine the effects of brief deprivation of auditory feedback on fricatives produced by cochlear implant users. The words /si/ and /ʃi/ were recorded by four children and four adults with their cochlear implant speech processor turned on or off. In the processor-off condition, word durations increased significantly for a majority of talkers. These increases were greater for children compared to adults, suggesting that children may rely on auditory feedback to a greater extent than adults. Significant differences in spectral measures of /ʃ/ were found between processor-on and processor-off conditions for two of the four children and for one of the four adults. These talkers also demonstrated a larger /s/-/ʃ/ contrast in centroid values compared to the other talkers within their respective groups. This finding may indicate that talkers who produce fine spectral distinctions are able to perceive these distinctions through their implants and to use this feedback to fine tune their speech. Two listening experiments provided evidence that some of the acoustic changes were perceptible to normal-hearing listeners. Taken together, these experiments indicate that for certain cochlear-implant users the brief absence of auditory feedback may lead to perceptible modifications in fricative consonants. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2167149]

PACS number(s): 43.70.Bk, 43.70.Ep, 43.71.Bp [AE] Pages: 1626–1635

I. INTRODUCTION

A growing body of evidence suggests that restoration of auditory feedback following cochlear implantation results in improvements in speech production by profoundly deaf individuals. Pre- versus post-implantation comparisons have revealed significant changes in speech over a span of weeks to months following the restoration of auditory feedback. These long-term changes include: shortened vowel durations, decreases in vowel F_0, lowered vowel intensity (SPL), shifts in the first and second formant frequencies of vowels toward more normal ranges, more appropriate voice onset times of stop consonants, and fricative spectral shifts toward more normal ranges (e.g., Tartter *et al.*, 1989; Economou *et al.*, 1992; Perkell *et al.*, 1992; Matthies *et al.*, 1994; Lane *et al.*, 1997; Kishon-Rabin *et al.*, 1999). These results confirm clinical findings that deaf individuals show marked improvements in a variety of speech attributes following restoration of auditory feedback. Perceptual data have indicated that these acoustic changes correspond with improvements in intelligibility and quality ratings. For example, normal-hearing judges perceived tokens produced by deaf persons after implantation to be qualitatively better than those produced before implantation (e.g., Tartter *et al.*, 1989; Kishon-Rabin *et al.*, 1999).

In addition to the long-term effects of auditory feedback, researchers also have explored the short-term effects of auditory feedback deprivation lasting anywhere from a few minutes to hours. These studies help address the minimal time window within which auditory feedback can have an effect on speech production. Most of these studies use a processor-on versus -off paradigm, in which an investigator can turn the cochlear-implant processor off to control the length of auditory deprivation and examine effects of complete absence of auditory feedback on speech production (e.g., Svirsky and Tobey, 1991; Svirsky *et al.*, 1992; Richardson *et al.*, 1993; Higgins *et al.*, 1999). In these studies where auditory deprivation lasted anywhere from 20 min–24 h, the following acoustic changes were noted: Formant frequency centralization for some vowels (e.g., /ɛ/ and /ɪ/), lengthened vowel durations, increases in SPL and F_0, and decreases in spectral contrast between /s/ versus /ʃ/. The findings of studies using the processor-on versus -off paradigm suggest that temporary changes in auditory feedback may result in significant modifications in both the segmental and suprasegmental properties of speech. These data are consistent with the idea that auditory feedback is used to calibrate or update the settings of the speech production mechanism. In other words, a cessation of auditory feedback presumably results in a loss of calibration and consequently a drift in speech parameters (e.g., Matthies *et al.*, 1996). To date, the effects of shorter periods of auditory deprivation (seconds to minutes) on speech production has not been thoroughly examined. Such information would be valuable because it would contribute to our understanding of the time window involved in calibration. Therefore, the first objective of this study was to investigate whether brief deprivation of auditory feedback (approximately 15–20 s) resulted in significant changes in the speech produced by cochlear implant users.

 0001-4966/2006/119(3)/1626/10/$22.50

A number of experiments with normal-hearing individuals also have investigated the short-term effects by examining modifications in the ongoing speech in response to altered auditory feedback (e.g., Kawahara and Williams, 1996; Houde and Jordan, 1998; Burnett *et al.*, 1998; Jones and Munhall, 2000). For instance, several studies have shown shifts in vowel fundamental frequency (F_0) within 100–150 ms following altered auditory feedback (e.g., Kawahara and Williams, 1996; Natke *et al.*, 2003). While some investigators have argued that these latencies would be too slow to assist in the online control of F_0 (e.g., Donath *et al.*, 2002; Natke *et al.*, 2003), others have suggested that auditory feedback may be used to regulate F_0 across the length of a syllable, depending on the naturalness of the speaking task (e.g., Xu *et al.*, 2004). Although the exact implications of these short-term shifts in F_0 remain debatable, most researchers agree that auditory feedback may be used to calibrate/update the settings of the "internal model" or representation in normal-hearing individuals (e.g., Kawahara and Williams, 1996; Jones and Munhall, 2000; Larson *et al.*, 2001).

The concept of an "internal model" has been elaborated in a recent theory that addresses the role of auditory feedback in speech production (Perkell *et al.*, 1997; Perkell *et al.*, 2000). According to this theory, speech sounds are produced with reference to an internal model that is acquired during childhood. This model is a mapping between the vocal tract configuration and its acoustic output and is learned with the help of auditory, somato-sensory, and visual feedback. With maturation, the internal model becomes more robust and feedback is used intermittently to calibrate or maintain the settings of the model. Thus, an absence of auditory feedback may result in a loss of calibration of the suprasegmental and segmental properties of speech and consequently lead to a drift in speech. If the internal model is dependent on feedback for its development and if it is also assumed that congenitally deafened children have less overall experience in shaping their internal model compared to postlinguistically deafened adults, then it may be predicted that briefly eliminating auditory feedback will produce greater changes in the speech of children than adults. Therefore, a second objective of this study was to compare feedback-related changes in the speech of deaf children and deaf adults.

While there is evidence for acoustic changes in speech production in the absence of auditory feedback, data on the perceptual consequences of short-term acoustic changes are limited. An investigation by Tye-Murray and colleagues (Tye-Murray *et al.* 1996), examined if changes that occurred between processor-on versus processor-off speaking conditions were perceptible to experienced listeners. Broad and narrow phonetic transcriptions of speech produced by 11 prelinguistically deafened children did not reveal any significant differences across the two speaking conditions in terms of place of articulation, height, or the voicing features of consonants and vowels. However, it is possible that phonetic transcription measures were not sensitive enough to capture the subtle acoustic variations that occurred in speech produced under altered auditory feedback conditions. Therefore, a final objective of this study was to utilize a quality rating scale measure to examine if normal-hearing judges can perceive acoustic changes in speech produced in the temporary absence of auditory feedback.

In summary, the objective of the present study was to investigate the short-term effects of auditory feedback on fricative-vowel words produced by adults and children in the presence and absence of auditory feedback delivered via cochlear implants. Fricative-vowel words were selected as stimuli for several reasons. First, sibilant sounds require relatively precise articulation with few obvious visual cues. As such, they may be considered prime targets for articulatory disruption under altered auditory feedback conditions. Second, /s/ and /ʃ/ have well-defined spectral properties that can be described and classified by spectral moments analysis (mean, variance, skewness, and kurtosis; Forrest *et al.*, 1988). Acoustic analyses and listening experiments were designed to: (a) examine effects of brief deprivation of auditory feedback on word durations and spectral measures of fricative-vowel words; (b) investigate group differences in measured parameters and examine whether children demonstrate larger speech changes than adults in the absence of auditory feedback; and (c) assess whether the auditory feedback-influenced acoustic changes are perceptible to normal-hearing adult listeners.

II. ACOUSTIC ANALYSES OF FRICATIVE-VOWEL WORDS

A. Methods

1. Participants

Participants included four postlinguistically deafened adults (A1–A4) and four congenitally deafened children (C1–C4). All were native, monolingual speakers of American English. Information concerning age, gender, etiology of deafness, age of implantation, and length of implant use are reported in Table I. To constrain potential variability associated with the use of different speech coding strategies, only participants using a Nucleus multichannel cochlear implant with the SPEAK speech coding strategy were included. All participants had used their implants for at least four years and had used the oral-aural mode of communication. Participants were paid for their participation.

2. Speech materials and procedures

Speech materials included the words /si/ and /ʃi/. Participants were seated comfortably in a sound-treated room. They were shown the words "see" and "she" written on index cards and were asked to produce each of these two words eight times in succession following a prompt by the examiner. The words were produced under two conditions: (a) with cochlear implant device turned on [processor-on] and (b) immediately after the device was turned off [processor-off]. In the processor-off condition, the device was turned off for approximately 15–20 s during the production of the eight tokens of "see" and "she." A total of 32 fricative-vowel words were elicited from each of the participants, yielding a set of 256 productions (2 consonants × 2 conditions × 8 repetitions × 8 talkers [4 adults and 4 children]). Speech samples were audio-recorded in a sound-treated room using

TABLE I. Demographic information for child and adult participants in the study.

Talker	Age (yrs)	Sex	Etiology	Implanted age (yrs)	Length of CI use (yrs)
C1	7:11	M	Congenital	2:4	5:7
C2	7:1	F	Congenital	2	5:1
C3	7	F	Congenital	2:1	4:11
C4	10:9	F	Congenital	5:5	5:5
A1	58	F	Meniere's disease	52	6:0
A2	46	F	Oto-sclereosis	39	7:0
A3	46	F	Pendrid syndrome	32	14:0
A4	41	F	Unknown	37	~4

an AIWA HD-X3000 digital audiotape (DAT) recorder with a Shure SM94 condenser microphone placed 10 in. from the speaker's mouth. Digital audio recordings were transferred to computer hard disk at a sampling rate of 48 kHz and a 16 -bit rate using a Townsend DAT Link+. Data were downsampled from 48 to 22 kHz for analyses. Spectral moments of fricative consonants and word durations (speaking rate) were derived from these recordings.

Centroid, variance, skewness, and kurtosis were estimated from spectral displays with a 20-ms full Hamming window centered at the fricative midpoint using TF32 software (Milenkovic, 2002). Spectral moments analysis treats the acoustic spectrum as a random probability distribution from which the four moments are derived: Centroid (spectral mean), variance (energy spread around the spectral peak), skewness (tiltedness or symmetry of the spectrum), and kurtosis (peakedness of the spectrum) (Forrest *et al.*, 1988; Tjaden and Turner, 1997). Word durations (beginning of the fricative to vowel offset) were measured using both waveform and spectral displays. The beginning of frication noise was defined as the first visible evidence of high-frequency energy preceding the vowel, while vowel offset was defined as the zero crossing of the last glottal pulse in the vocalic portion of the syllable.

B. Results

Cochlear implant users have different histories of profound deafness and psychophysical capabilities due to different patterns of nerve survival. For these reasons, individuals who have cochlear implants may be expected to respond differently to the absence of stimulation (Svirsky *et al.*, 1992). To examine these individual talker characteristics, data for each talker were analyzed separately. Group comparisons were also made to address the question of whether children as a group differed from adult talkers in their responses to deprivation of auditory feedback. The primary aim of the group analyses was to test the hypothesis that the magnitude of changes in speech patterns with and without feedback will be larger in children compared to adults. Therefore, a difference measure ($\triangle$, the absolute value of processor off minus on) was used in group comparisons. Effect size ("*d*"), a measure of the magnitude of treatment effect is reported along with the rest of the statistics for both individual and group data (see Cohen, 1988).

1. Individual subject data

A three-factor [Condition × Consonant × Trial] analysis of variance (ANOVA) was conducted separately for each talker for the following variables: spectral moments (centroid, skewness, variance, and kurtosis) and word durations. Two planned comparisons were conducted (with *Bonferroni* corrections and a family-wise p value set at $p < 0.025$) to test the hypotheses that acoustic measures for /si/ and /ʃi/ were significantly different in processor-on versus processor-off conditions.

a. Fricative spectral moments measures *Centroid*: As shown in Fig. 1 (left panel), centroid values were significantly higher in processor-on versus -off condition for /ʃ/ produced by talker C1 [$F(1,7)=20.43$; $p=0.002$; $d=1.76$]. In addition, centroid values were significantly lower in

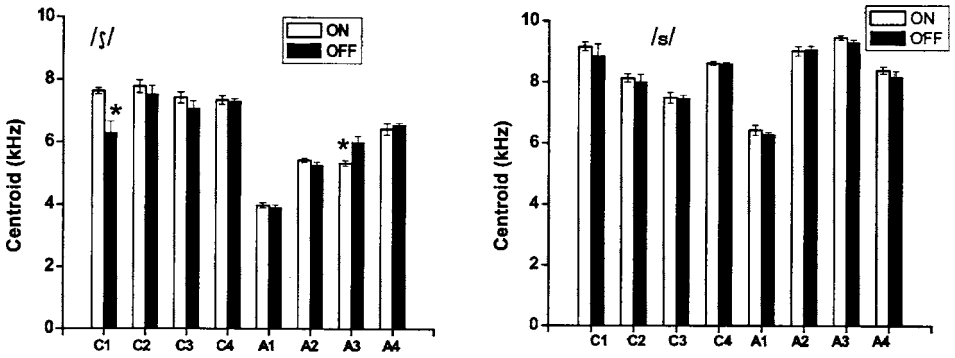


FIG. 1. Average centroid values of fricatives /ʃ/ and /s/ produced by deaf children (C1–C4) and deaf adults (A1–A4) in processor-on and -off conditions. Error bars represent the standard error across eight repetitions. A star [*] indicates significant difference ($p < 0.025$).

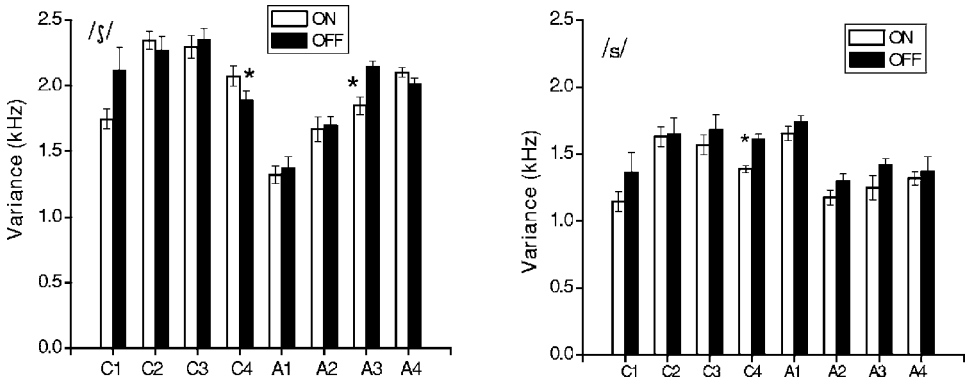


FIG. 2. Average variance values of fricatives /ʃ/ and /s/ produced by deaf children (C1–C4) and deaf adults (A1–A4) in processor-on and -off conditions. Error bars represent the standard error across eight repetitions. A star [*] indicates significant difference ($p<0.025$).

processor-on versus -off conditions for /ʃ/ produced by talker A3 [$F(1,7)=9.23$; $p=0.0006$; $d=1.63$]. No reliable differences were observed for the remaining talkers. As shown in the right panel, there were no significant differences between processor conditions for consonant /s/ for any of the talkers.

Variance: Figure 2 shows variance measures in processor-on versus processor-off conditions for all talkers. As shown in the left panel, talkers C4 [$F(1,7)=17.15$; $p=0.02$; $d=0.94$] and A3 [$F(1,7)=10.26$; $p=0.003$; $d=1.96$] demonstrated significant differences for /ʃ/ across the processor conditions. While C4 showed a significant increase in variance values, talker A3 showed a decrease in variance values in processor-on compared to -off conditions. Also, as illustrated in the right panel, significant decreases were found for /s/ produced in processor-on compared to -off conditions by talker C4 [$F(1,7)=11.53$; $p=0.009$; $d=2.5$]. No reliable differences were observed for the other talkers.

Skewness: Significant changes were noted for /ʃ/ for C1 [$F(1,7)=9.51$; $p=0.01$; $d=1.72$], and A3 [$F(1,7)=9.08$; $p=0.003$; $d=1.40$] in processor-on versus processor-off conditions [see Fig. 3, left panel]. C1 showed negative skewness coefficients in processor-on and positive values in processor-off conditions. On the other hand, A3 showed higher, positive skewness coefficient values in processor-on compared to -off conditions. No reliable differences were observed for the other talkers. For /s/, none of the talkers showed any significant differences between processor conditions [see right panel]. Skewness coefficients were negative for /s/ in a majority of talkers, except talker A1. As shown in the Fig. 3 (right panel), talker A1 showed positive skewness coefficients in both processor-on and -off conditions. It should be noted that these values are typical for /ʃ/ productions.

Kurtosis: Talkers A1 and A3 showed a relatively large processor-on versus -off differences for /ʃ/ compared to the rest of the talkers. Similarly, C4, A2, and A3 showed relatively large processor-on versus -off differences for /s/ compared to the rest of the talkers. However, these differences should be considered as trends, as they did not reach statistical significance.

b. Word durations Previous studies have reported no differences in mean duration of /s/ versus /ʃ/ and mean duration of /sVC/ versus /ʃVC/ syllables (e.g., Jongman *et al.*, 2000) and none were found in the present study. Therefore, duration data were collapsed across the two words, /si/ and /ʃi/. As shown in Fig. 4, turning the speech processor off resulted in significantly longer word durations for all four children and was confirmed by a significant main effect of condition: C1 [$F(1,7)=15.80$; $p=0.0054$; $d=0.62$]; C2 [$F(1,7)=6.76$; $p=0.03$; $d=1.39$]; C3 [$F(1,7)=10.76$; $p=0.01$; $d=0.86$]; and C4 [$F(1,7)=7.09$; $p=0.03$; $d=1.13$]. Two adults (A2 and A3) also lengthened word durations in the processor-off condition: A2 [$F(1,7)=16.98$; $p=0.004$; $d=1.70$], and A3 [$F(1,7)=6.26$; $p=0.04$; $d=0.64$]. In contrast, talker A4 demonstrated significantly shorter word durations in processor-off condition [$F(1,7)=5.86$; $p=0.04$; $d=1.00$], while A1 did not show significant changes in word durations as a function of processor condition. Increases in both fricative and vowel durations contributed to prolonged word durations following cessation of auditory feedback.

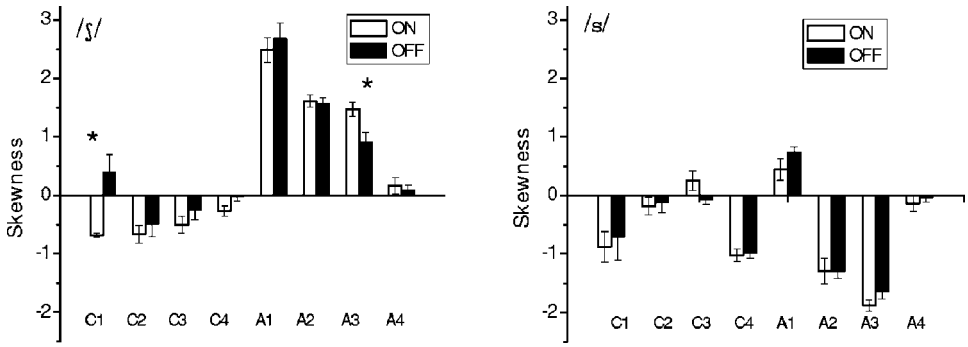


FIG. 3. Average skewness values of fricatives /ʃ/ and /s/ produced by deaf children (C1–C4) and deaf adults (A1–A4) in processor-on and -off conditions. Error bars represent the standard error across eight repetitions. A star [*] indicates significant difference ($p<0.025$).

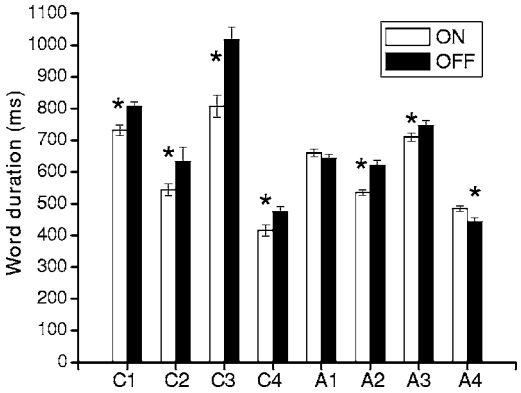


FIG. 4. Average word durations in processor-on and -off conditions for words produced by deaf adults and children. Error bars represent the standard error across eight repetitions of words /si/ and /ʃi/. A star [*] indicates significant difference ($p < 0.025$).

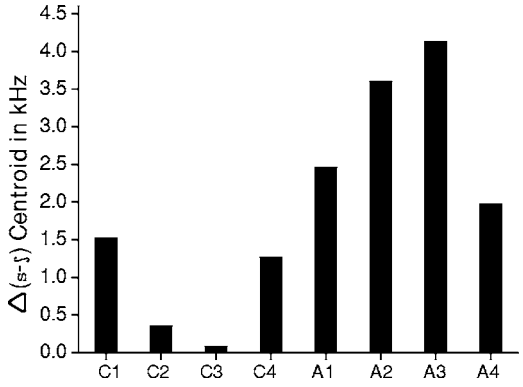


FIG. 5. Average /s/-/ʃ/ (△) centroid values for four deaf children (C1–C4) and four deaf adults (A1–A4) in processor-on conditions.

2. Congenitally deafened children versus postlinguistically deafened adults

A subset of children (C1, C4) and adults (A3) showed significant changes in spectral moments while a majority of the participants showed significant changes in word duration in processor-on versus -off conditions. Therefore, analyses examining whether the magnitude of processor-on versus processor-off differences is greater in children's compared to adults' speech were restricted to word duration data. A second series of group comparisons examined whether children and adults differ in production of /s/-/ʃ/ contrast in centroid values in the processor-on condition.

a. Fricative spectral moments measures Table II shows averaged spectral moments measures for /s/ and /ʃ/ produced by congenitally deafened children and postlinguistically deafened adults. The spectral moment measures derived for postlinguistically deafened adults were generally comparable to ranges reported for normal-hearing adults: Centroid (Nittrouer, 1995); variance, skewness (/s/ only), and kurtosis coefficients (Jongman *et al.*, 2000). Similarly, spectral moment measures derived for congenitally deafened children were comparable to ranges reported for normal-hearing children: Centroid (Pentz *et al.*, 1979; Nittrouer, 1995), skewness coefficients, and kurtosis coefficients for /ʃ/ only (Uchanski and Geers, 2003).

As shown in Table II, congentially deafened children showed higher centroid values and greater variability for /ʃ/ than postlinguistically deafened adults. Further, the variance values for /ʃ/ also were higher for children compared to adults. These findings may suggest developmental differences in /ʃ/ productions.

Figure 5 shows /s-ʃ/ contrasts in processor-on condition for each of the child and adult talkers. From this figure, it is clear that talkers C1 and C4 produced the largest /s/-/ʃ/ contrast among the children, while A3 produced largest /s/-/ʃ/ contrast among the adults. These three talkers also showed significant changes in fricative spectral moments in processor-on versus -off conditions. A closer inspection of Table II also shows that in the processor-on condition, adults produced an average /s/-/ʃ/ centroid difference values of approximately 3 kHz, while children produced a /s/-/ʃ/ centroid

TABLE II. Mean (*M*) and standard deviations (*SD*) of spectral moments for fricatives produced by congenitally deafened children (C1–C4) and postlinguistically deafened adults (A1–A4) in processor-on and processor-off conditions. Values for centroids and variance are reported in kilohertz. Skewness and kurtosis coefficients are dimensionless.

		Centroid				Variance				Skewness				Kurtosis			
		/s/		/ʃ/		/s/		/ʃ/		/s/		/ʃ/		/s/		/ʃ/	
		ON	OFF	ON	OFF	ON	OFF	ON	OFF	ON	OFF	ON	OFF	ON	OFF	ON	OFF
C1	*M*	9.15	8.84	7.62	6.29	1.14	1.36	1.74	2.11	−0.87	−0.70	−0.68	0.39	1.33	1.33	0.07	−0.15
	SD	0.41	1.10	0.28	1.03	0.20	0.418	0.21	0.49	0.73	1.12	0.10	0.87	1.71	2.46	0.59	1.25
C2	*M*	8.12	7.99	7.77	7.52	1.62	1.65	2.34	2.27	−0.17	−0.12	−0.65	−0.49	−0.80	−0.52	−0.81	−0.62
	SD	0.40	0.73	0.58	0.76	0.20	0.34	0.19	0.28	0.41	0.46	0.42	0.59	0.46	1.24	0.48	0.83
C3	*M*	7.49	7.47	7.41	7.07	1.56	1.67	2.29	2.34	0.25	−0.06	−0.49	−0.24	−0.004	0.17	−0.87	−1.14
	SD	0.50	0.30	0.50	0.68	0.20	0.32	0.24	0.25	0.48	0.23	0.40	0.46	0.98	0.94	0.78	0.37
C4	*M*	8.61	8.58	7.33	7.28	1.38	1.61	2.07	1.88	−1.01	−0.98	−0.26	−0.02	1.18	0.46	−0.81	−0.73
	SD	0.17	0.15	0.40	0.28	0.07	0.10	0.21	0.19	0.29	0.23	0.25	0.22	1.18	0.80	0.33	0.27
A1	*M*	6.43	6.29	3.97	3.89	1.65	1.73	1.32	1.37	0.44	0.74	2.48	2.68	0.11	−0.02	7.02	8.21
	SD	0.48	0.20	0.22	0.23	0.14	0.14	0.18	0.23	0.52	0.23	0.60	0.73	0.72	0.57	4.16	4.77
A2	*M*	9.02	9.06	5.41	5.25	1.17	1.29	1.66	1.69	−1.28	−1.28	1.61	1.56	3.97	2.96	2.01	1.83
	SD	0.44	0.34	0.18	0.30	0.15	0.16	0.26	0.18	0.61	0.34	0.31	0.29	2.43	1.43	1.59	1.30
A3	*M*	9.46	9.30	5.32	5.99	1.24	1.41	1.84	2.14	−1.88	−1.64	1.47	0.90	6.05	4.02	1.29	−0.48
	SD	0.19	0.25	0.24	0.55	0.25	0.13	0.18	0.12	0.25	0.35	0.33	0.47	1.97	1.67	1.22	0.96
A4	*M*	8.39	8.17	6.41	6.53	1.31	1.36	2.10	2.01	−0.13	−0.03	0.16	0.08	−0.31	−0.48	−1.29	−1.29
	SD	0.32	0.55	0.52	0.23	0.13	0.30	0.10	0.12	0.39	0.21	0.37	0.25	0.42	0.85	0.160	0.18

difference of approximately 800 Hz. Similarly, in the processor-off condition, adults produced an average contrast in centroid values of approximately 2.78 kHz between /s/ and /ʃ/, while children produced an average contrast of approximately 1.17 kHz. Analyses examining group differences for /s/-/ʃ/ contrast indicated that children produced a significantly less distinct /s/-/ʃ/ contrast in comparison to adults for both processor-on [$F(1,7)=91.09$; $p=0.0001$; $d=2.41$] and processor-off conditions [$F(1,7)=47.32$; $p=0.0001$; $d=1.4$].

b. Word duration Group comparisons were conducted to examine whether children showed larger processor-on versus -off differences in word duration using a one-way ANOVA. Note that processor on minus off difference values [△] were used in these analyses. Previous studies have reported no differences in mean duration of /s/ versus /ʃ/ and mean duration of /sVC/ versus /ʃVC/ syllables (e.g., Jongman, *et al.* 2000). Therefore, duration data were collapsed across the two words, /si/ and /ʃi/.

Results indicated larger processor-on versus processor-off differences for word durations produced by children [Mean (*M*): 169.75 ms; Standard deviation (*SD*): 148.35] compared to adults (*M*: 78.48 ms; *SD*: 53.52) and was confirmed by a significant main effect of group [$F(1,7)=79.51$; $p=0.0001$; $d=0.81$]. However, word durations are generally longer in children's compared to adults' speech. Therefore, to account for these relative differences, the △ word duration values also were adjusted by computing the proportion of change in each talker group [defined as (on-off)/(on) expressed as absolute values]. The proportion of change in processor-on versus -off conditions was 0.27 and 0.13 for children and adults, respectively. This finding confirms that the proportion of change in word duration was, in fact, longer for children's words compared to those of adults, and was not merely a by-product of the relatively longer duration of children's utterances.

C. Summary

Auditory deprivation lasting for approximately 15–20 s resulted in significant spectral changes in /ʃ/ produced by two children (C1, C4) and one adult (A3). These talkers also demonstrated a larger /s/-/ʃ/ contrast in centroid values, compared to the other talkers in their group. Further, results showed significant changes in word durations in the absence of auditory feedback for a majority of talkers. Group comparisons revealed a significantly larger change in word durations in processor-on versus -off conditions for children compared to adults. In addition, children produced a significantly less distinct /s/-/ʃ/ contrast compared to adults. A quality rating study was conducted to examine the perceptual relevance of these acoustic findings.

III. RATING EXPERIMENT

The objective of this experiment was to determine whether the acoustic changes that resulted from changes in auditory feedback status were perceptible to normal-hearing listeners. Speech samples produced by congenitally deafened children and postlinguistically deafened adults were presented to normal-hearing listeners to determine whether the acoustic changes that occurred in the absence of auditory feedback were detectable. Based on the acoustic findings of developmental differences for /ʃ/, and the finding of processor-on versus -off differences for /ʃ/ produced by children, it was predicted that quality ratings provided by listeners would be significantly poorer for /ʃ/ produced by children in processor-off compared to -on conditions.

A. Methods

1. Listeners

Ten native speakers of American English with a background in Speech-Language-Pathology served as listeners. All listeners had taken a graduate-level course in phonetics. None reported any speech or hearing problems.

2. Stimuli

Stimuli for this experiment included the fricatives /s/ and /ʃ/ excised from the words /si/ and /ʃi/ used in the acoustic analyses. The beginning of each fricative was identified by the onset of high frequency noise components, and the fricative end was identified as the intensity minimum immediately preceding the onset of vowel periodicity (Jongman *et al.* 2000). A total of 256 stimuli [8 talkers × 2 consonants × 2 conditions × 8 repetitions] were included in the rating experiment. Fricative noises excised from the fricative-vowel words were included as stimuli to minimize any potential influences of the following vowel during fricative judgments.

3. Procedure

Listeners were asked to assign quality ratings on a 100 mm Visual Analog Scale (VAS) displayed on the computer monitor. The VAS was used to gather perceptual data since it is frequently used in the evaluation of pathological voices and has been shown to correlate well with instrumental measures (e.g., Rabinov *et al.*, 1995; Yu *et al.*, 2002).

Listeners were informed that they would hear fricatives produced by cochlear-implanted children and adults under different talking conditions. They were asked to attend to the overall quality of the fricative noise and to provide ratings on a poor-good quality continuum based on how well the fricative represented the intended phonetic category. Sounds were presented diotically while the intended target ("s" or "sh") was displayed on the monitor. The listeners assigned ratings for fricative productions by moving a slider on a 100 mm Visual Analog Scale on the computer monitor to a desired point on a poor-good quality continuum. Listeners used a computer mouse to position the slider. The experiment was self-paced and listeners were allowed to listen to stimuli any number of times before assigning a rating by pressing the replay button. The experiment lasted approximately 45 min.

B. Results

A three-way repeated measures ANOVA was performed with condition (ON, OFF), consonant (/s/, /ʃ/), and group (children, adults) as within-subject factors. Results showed that listeners assigned higher ratings for fricative consonants produced in processor-on than in processor-off conditions (*M*: 63.09; *SD*: 16.55 versus *M*: 57.86; *SD*: 20.43), and was confirmed by a significant main effect of condition [$F(1,9)$

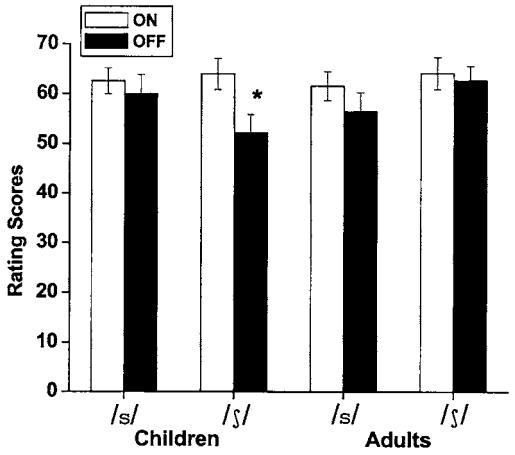


FIG. 6. Average rating scores assigned by ten listeners for fricative consonants /ʃ/ and /s/ spoken by deaf children (C1–C4) and deaf adults (A1–A4) in processor-on and -off conditions. A star [*] indicates significant differences in rating scores ($p<0.0125$).

$=57.48$; $p<0.05$]. This finding indicates that listeners are sensitive to speech production differences due to processor-on versus -off conditions. Analyses also revealed a significant three-way interaction: Group × condition × consonant [$F(1,9)=19.32$; $p<0.05$]. Figure 6 shows rating scores assigned by the ten listeners for /s/ and /ʃ/ produced by children and adults in processor-on and processor-off conditions. *Bonferroni* corrections for multiple comparisons (with family-wise p value set at $p<0.0125$) confirmed that listeners assigned a significantly higher rating for the consonant /ʃ/ (M: 64; SD: 17.61 versus M: 52.16; SD: 21.13) produced by children in the processor-on compared to the processor-off condition [$F(1,9)=64.53$; $p=0.0001$; $d=0.37$].

Overall, the results correspond well with the acoustic findings of significant changes in the consonant /ʃ/ produced by children in the processor-on compared to -off conditions. However, informal comments by some of the participants revealed a potential problem with this experiment, namely, that excised fricatives sound unnatural and may not always be perceived as speech. Therefore, an identification experiment was conducted to rule out this possibility and to assess the intelligibility of fricative-vowel words produced in both processor-on and -off conditions.

IV. IDENTIFICATION EXPERIMENT

The purpose of this experiment was to determine the identification accuracy of fricative-vowel words in both processor-on and processor-off conditions. Words produced by congenitally deafened children and postlinguistically deafened adults were presented to normal-hearing listeners to determine how well they could identify fricative-vowel words spoken in processor-on versus processor-off speaking conditions.

A. Methods

1. Listeners

Ten listeners participated in this experiment. These listeners were different from those who participated in the quality rating experiment. Participants were native speakers of American English with a background in Speech-Language-Pathology. All listeners had taken a graduate-level course in phonetics and none reported speech or hearing problems.

2. Stimuli

Stimuli for the identification experiment included the fricative-vowel words /si/ and /ʃi/ used in the acoustic analyses. Whole words rather than excised fricative consonants were used in this experiment to avoid the use of truncated speech segments which may sound unnatural. A total of 256 stimuli (8 talkers × 2 words × 2 conditions × 8 repetitions) were used in this experiment.

3. Procedure

Listeners heard words /si/ and /ʃi/ produced by cochlear-implanted children and adults. The task was to identify what they heard, by clicking on one of the response boxes on a computer monitor labeled "see" or "she." The experiment lasted approximately 15 min.

B. Results

A three-way repeated measures ANOVA was performed with condition (ON, OFF), consonant (/s/, /ʃ/), and group (children, adults) as within-subject factors. In contrast to the rating experiment, results did not reveal a significant main effect for condition. However, similar to the rating experiment, analyses revealed a significant three-way interaction: Group × condition × consonant [$F(1,9)=16.45$; $p=0.0001$]. Planned comparisons with *Bonferroni* corrections (family-wise p value set at $p<0.0125$) confirmed that identification accuracy was significantly higher for /ʃ/ (M: 98.43; SD: 3.68 versus M: 91.25; SD: 16.21) produced by children in processor-on compared to -off conditions [$F(1,9)=11.04$; $p=0.0001$; $d=0.34$].

Figure 7 shows identification scores for words (/si/ and /ʃi/) produced by each of the eight talkers in processor-on and -off conditions. There were fairly strong ceiling effects, with the identification scores for the words /si/ and /ʃi/ produced by all talkers in the processor-on condition in the range of 90–100%. These high scores confirm that a majority of the fricative-vowel words were accurate productions and that they were highly intelligible. To examine the difference in identification accuracy for these words in processor-on versus -off conditions, identification data were analyzed separately for each of the talkers. A two-way repeated measures ANOVA was performed with condition (ON, OFF), and words (/si/, /ʃi/) as within-subject factors. Planned comparisons with *Bonferroni* corrections (family-wise p value set at $p<0.025$) were performed to test the prediction that the identification accuracy was significantly different for words /si/ and /ʃi/ produced in processor-on versus processor-off conditions. As shown in Fig. 7, the identification accuracy

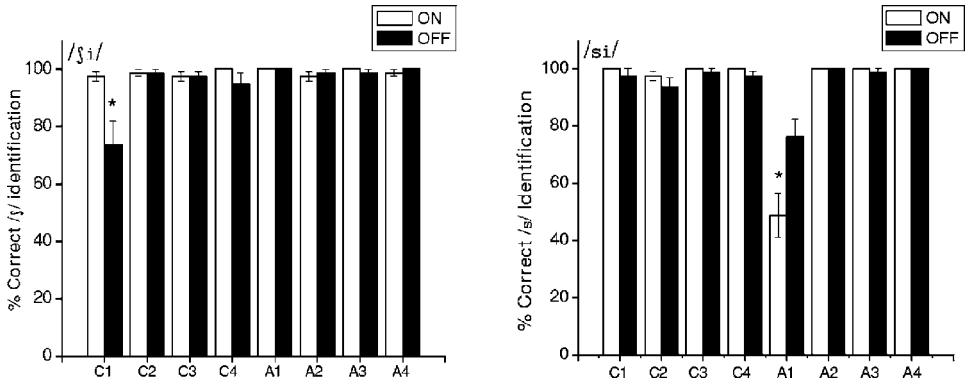


FIG. 7. Average identification accuracy scores based on listening judgments provided by ten normal-hearing listeners for words /ʃi/ and /si/ spoken by deaf children (C1–C4) and deaf adults (A1–A4) in processor-on and -off conditions. A star [*] indicates significant differences in identification scores ($p < 0.025$).

for /ʃi/ produced by talker C1 was significantly higher in processor-on compared to processor-off conditions [$F(1,9) = 18$; $p = 0.0001$; $d = 0.72$]. Identification accuracy was higher for /ʃi/ produced by C4 in processor-on compared to processor-off conditions, however, these values failed to reach significance. Surprisingly, analyses also showed significantly lower identification accuracy for /si/ produced by A1 [$F(1,9) = 31.57$; $p = 0.0001$; $d = 0.59$] in processor-on compared to processor-off conditions. No reliable differences in processor-on versus -off conditions were found for rest of the talkers.

C. Perception Experiments Summary

The main results of the rating experiment were significantly higher ratings in processor-on versus -off conditions and significantly higher ratings for consonant /ʃ/ produced by children in processor-on compared to -off conditions. Results of the identification experiment confirmed this finding, by showing higher identification accuracy for /ʃ/ produced by children in processor-on compared to -off conditions. In addition, data from the identification experiment showed higher accuracy for /ʃ/ produced by talker C1 in processor-on versus -off conditions.

V. GENERAL DISCUSSION

Fricative spectral moments and word durations were measured from speech samples produced by adults and children with and without auditory feedback provided by a cochlear implant device. In the brief absence of auditory feedback (approximately 15–20 s), significant changes in fricative spectral measures were noted for a subset of talkers (C1, C4, and A3). Data further revealed that these processor-on versus processor-off differences in the spectral measures were present primarily for /ʃ/, rather than /s/. These data are similar to other studies which note significant spectral differences for /ʃ/ in processor-on versus processor-off conditions (Matthies *et al.* 1996).

There are at least two possible explanations for why significant changes were found for /ʃ/, and not /s/. One reason may be that the alveolar ridge provides an anchor point for the articulation of /s/ in the absence of auditory feedback. There is no such well-defined landmark for articulation of /ʃ/. Thus, the articulation of /ʃ/ is likely to drift in the absence of auditory feedback (Matthies *et al.* 1996). Another explanation may involve saturation effects. It has been documented that /s/ is characterized by strong saturation effects (quantal effects) compared to /ʃ/ (see Perkell *et al.* 2004). In other words, the production of /s/ results in stable acoustic cues, despite variability in articulation in the absence of feedback. Thus, in the absence of auditory feedback, the production of /ʃ/ is more likely to drift than /s/.

In the brief absence of auditory feedback, significant changes also were noted in word durations (speaking rate) for seven of the eight talkers. Of these seven talkers, six showed increases and one showed a decrease in word duration during the processor-off condition. These results confirm the prediction that suprasegmental properties are influenced by temporary changes in auditory feedback status to a greater extent than segmental aspects of speech (Perkell *et al.*, 1997; Perkell *et al.*, 2000). The finding that feedback-related changes were observed both for adults and children also replicates the findings of Tobey *et al.* (2000), who reported that syllable durations were significantly shorter in processor-on versus -off conditions in young cochlear-implanted children. The increases in word duration in the absence of auditory feedback are most likely a result of the adaptive strategies used by deaf speakers to maximize feedback via proprioceptive and tactile channels (e.g., Svirsky *et al.*, 1992; Higgins *et al.*, 1999). The fact that talker A4 showed a decrease in word duration in the absence of auditory feedback might suggest individual differences in the use of adaptive strategies by deaf talkers. Similar results have been reported by several other studies that have attributed these individual variations in speech production to differences in speech intelligibility, speech perception ability, etiology of deafness, electrode insertion depth and viability of neural population (Svirsky *et al.*, 1992; Perkell *et al.*, 1992; Richardson *et al.*, 1993; Tye-Murray *et al.*, 1996).

Overall, the present duration data demonstrated that in the brief absence of auditory feedback, a majority of talkers decreased their speaking rate. However, there was a fair amount of individual talker variability with respect to changes in fricative spectral measures resulting from a change in hearing status. This individual talker variability must be explained as it addresses the issue of speech percep-

tion and production links. As expected, C1 lengthened word durations in the processor-off condition. In addition, C1 showed a greater /s/-/ʃ/ contrast in the absence of auditory feedback. These findings are consistent with the idea that some talkers attempt to produce clear speech when auditory feedback is impoverished (e.g., Krause and Braida, 2004).

Only a subset of talkers (C1, C4, and A3) showed fricative spectral changes in the absence of auditory feedback. Why did these talkers demonstrate these changes and not others? Perkell and colleagues (Perkell *et al.*, 2004) noted that talkers who produce a distinct sibilant contrast are also adept at perceiving fine acoustic differences between /s/ and /ʃ/. In the present data, A3 produced the greatest sibilant contrast among the adults, while C1 and C4 produced the greatest sibilant contrast among the children (see Fig. 5). It is possible that these talkers were able to perceive fine acoustic differences in the sibilants through their implants and use the feedback information to fine tune their speech. Analogously, talkers who rely on auditory feedback to fine tune their speech may demonstrate modifications in speech when it is absent or altered. These preliminary findings emphasize the linkage between speech perception and production, which merits careful examination in future studies.

It was predicted that congenitally deafened children would show larger processor-on versus -off differences than postlinguistically deafened adults. This expectation was confirmed in the word duration data. Children demonstrated greater processor-on versus -off differences in word duration than adults. Group comparisons also revealed differences in the production of /ʃ/, but not for /s/, suggesting that some of the children are continuing to refine /ʃ/ production.

Further analyses revealed that postlinguistically deafened adults, as a group, produced a significantly larger /s/-/ʃ/ contrast than congenitally deafened children. This finding indicates a developmental difference in fricative production and further suggests that 7–11 year old congenitally deafened children, with 3–5 years of implant experience, are continuing to refine their fricative production, for which auditory feedback delivered through their cochlear implants may be crucial.

Two listening experiments were conducted to examine whether the acoustic changes that occurred in the absence of auditory feedback were perceptible to normal-hearing listeners. Results from the rating experiment showed lower ratings for processor-off compared to -on conditions, suggesting that VAS can be a sensitive tool to assess speech production changes in the absence of auditory feedback. As expected, results also showed lower ratings for /ʃ/ produced by children in processor-off compared to -on conditions. These data corroborate the findings of the acoustic analyses and further suggest that some congenitally deafened children produced perceptible acoustic changes in fricatives when speaking in the absence of auditory feedback.

A second listening experiment was designed to examine the overall intelligibility of the fricative-vowel words produced in processor-on and -off conditions. Results showed that the identification accuracy for words produced by all eight talkers ranged from 90–100 %, suggesting that the fricative consonants were well produced. Based on the acoustic analyses, it was expected that the identification accuracy for talkers C1, C4, and A3 would be different in processor-on versus -off conditions. Consistent with the predictions, identification accuracy for both C1 and C4 were higher in processor-on versus -off conditions, although the differences were significant only for C1. Talker A3 produced significant acoustic changes in /ʃ/ in processor-on versus -off conditions. However, this talker also demonstrated the largest /s/-/ʃ/ absolute centroid difference of her talker group. Thus, for this talker, the processor-on versus -off acoustic differences while appreciable, may not have been sufficient to result in a perceptual change for the listeners. Identification accuracy for A1 was significantly different for /s/ in processor-on versus -off conditions. The skewness coefficients for /s/ produced by A1 were more in the range of /ʃ/ in both processor-on and -off conditions. It is possible that listeners were sensitive to variations in skewness coefficient, leading to confusions and errors in judgment for /s/ produced in both processor-on versus -off conditions. The identification data also corroborated some of the findings of the rating experiment by showing lower identification accuracy for /ʃ/ produced by children in processor-off compared to -on conditions. Taken together, the rating and identification experiments suggested that some deaf individuals may produce perceptible acoustic changes in their speech as a result of brief auditory feedback deprivation.

In summary, in the brief absence (approximately 15–20 s) of auditory feedback significant changes occurred in word duration for a majority of talkers. In addition, perceptible spectral changes were demonstrated in fricative consonants produced by a subset of talkers. There are at least three possible explanations for why only some talkers showed modifications in speech when auditory feedback was briefly eliminated. First, the brief lack of auditory feedback alters the parameter settings of the internal model, leading to modifications in speech. If so, this would fit well with the predictions in the literature that the internal model for both segmental and suprasegmental aspects of speech (e.g., Perkell *et al.*, 2000; Jones and Munhall, 2000) is responsive to changes in the acoustic environment. Second, some talkers may deliberately use a different speaking strategy in the absence of auditory feedback. In other words, some deaf talkers may attempt to speak clearly to maximize feedback via tactile and proprioceptive channels (e.g., Higgins *et al.*, 1999; see also Krause and Braida, 2004). Third, individuals with deafness may have two internal models; one acquired prior to cochlear implantation and another acquired after cochlear implantation.[1] The fact that some talkers demonstrated immediate modifications in speech when auditory feedback was removed might reflect a switch from the use of the current internal model to the one acquired prior to implantation. These different accounts should be examined carefully in future studies. Our understanding of the feedback and feedforward systems for speech production is evolving. Future experiments comparing the effects of short-term changes in auditory feedback (spanning milliseconds to seconds) on a variety of segmental and suprasegmental measures of speech

produced by individuals with normal hearing and deafness may be able to address the precise nature of the feedback and feedforward systems for speech production.

ACKNOWLEDGMENTS

This work was conducted in partial fulfillment of doctoral degree at the University of Texas at Dallas. Portions of this research were presented at the 143rd meeting of the Acoustical Society of America, Pittsburgh, PA, 2002. This work was supported by predoctoral fellowship awarded to the first author from the National Institutes of Health (NIDCD No. F31 DC05280-01, 2001). The authors thank all the participants and their families. Thanks to an anonymous reviewer for comments on an earlier version of the manuscript. Thanks also to Nils Penard for his input on statistical analyses and Megha Sundara for her comments on this manuscript.

[1]The authors thank Dr. Anders Lofqvist for this suggestion.

Burnett, T. A., Freedland, M. B., Larson, C. R., and Hain, T. C. (**1998**). "Voice f_0 responses to manipulations in pitch feedback," J. Acoust. Soc. Am. **103**, 3153–3161.

Cohen, J. (**1988**). *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum, Hillsdale).

Donath, T. M., Natke, U., and Kalveram, K. T. (**2002**). "Effects of frequency-shifted auditory feedback on voice F0 contours in syllables," J. Acoust. Soc. Am. **111**, 357–366.

Economou, A., Tartter, V. C., Chute, P. M., and Hellman, S. A. (**1992**). "Speech changes following re-implantation from a single-channel to a multichannel cochlear implant," J. Acoust. Soc. Am. **92**, 1310–1323.

Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. N. (**1988**). "Statistical analysis of word-initial voiceless obstruents: preliminary data," J. Acoust. Soc. Am. **84**, 115–123.

Higgins, M. B., McCleary, E. A., and Schulte, L. (**1999**). "Altered phonatory physiology with short-term deactivation of children's cochlear implants," Ear Hear. **20**, 426–438.

Houde, J. F., and Jordan, M. I. (**1998**). "Sensorimotor adaptation in speech production," Science **279**, 1213–1216.

Jones, J. A., and Munhall, K. G. (**2000**). "Perceptual calibration of F0 production: Evidence from feedback perturbation," J. Acoust. Soc. Am. **108**, 1246–1251.

Jongman, A., Wayland, R., and Wong, S. (**2000**). "Acoustic characteristics of English fricatives," J. Acoust. Soc. Am. **108**, 1252–1263.

Kawahara, H., and Williams, J. C. (**1996**). "Effects of auditory feedback on voice pitch trajectories: Characteristic responses to pitch perturbations," in *Vocal Fold Physiology: controlling complexity and Chaos*, edited by P. J. Davis and N. H. Fletcher (Singular, San Diego), pp. 263–278.

Kishon-Rabin, L., Taitelbaum, R., Tobin, Y., and Hildesheimer, M. (**1999**). "The effect of partially restored hearing on speech production of postlingually deafened adults with multichannel cochlear implants," J. Acoust. Soc. Am. **106**, 2843–2857.

Krause, J. C., and Braida, L. D. (**2004**). "Acoustic properties of naturally produced clear speech at normal speaking rates," J. Acoust. Soc. Am. **115**, 362–378.

Lane, H., Wozniak, J., Matthies, M., Svirsky, M., Perkell, J., O'Connell, M., and Manzella, J. (**1997**). "Changes in sound pressure and fundamental frequency contours following changes in hearing status," J. Acoust. Soc. Am. **101**, 2244–2252.

Larson, C. R., Burnett, T. A., Bauer, J. J., and Kiran, S. (**2001**). "Comparison of voice F0 responses to pitch-shift onset and offset conditions," J. Acoust. Soc. Am. **110**, 2845–2848.

Matthies, M. L., Svirsky, M., Lane, H., and Perkell, J. (**1994**). "A preliminary study of effects of cochlear implants on the production of sibilants," J. Acoust. Soc. Am. **96**, 1367–1373.

Matthies, M. L., Svirsky, M., Perkell, J., and Lane, H. (**1996**). "Acoustic and articulatory measures of sibilant production with and without auditory feedback from a cochlear implant," J. Speech Lang. Hear. Res. **39**, 936–946.

Milenkovic, P. (**2002**). *Time-Frequency Analyzer (TF32)*. Software analyses package.

Natke, U., Donath, T. M., and Kalveram, K. T. (**2003**). "Control of voice fundamental frequency in speaking versus singing," J. Acoust. Soc. Am. **113**, 1587–1593.

Nittrouer, S. (**1995**). "Children learn separate aspects of speech production at different rates: Evidence from spectral moments," J. Acoust. Soc. Am. **97**, 520–530.

Pentz, A., Gilbert, H. R., and Zawadzki, P. (**1979**). "Spectral properties of fricative consonants in children," J. Acoust. Soc. Am. **66**, 1891–1893.

Perkell, J. S., Lane, H., Svirsky, M. A., and Webster, J. (**1992**). "Speech of cochlear implant patients: A longitudinal study of vowel production," J. Acoust. Soc. Am. **91**, 2961–2979.

Perkell, J., Guenther, F. H., Lane, H., Matthies, M., Perrier, P., Vick, J., Wilhelms-Tricarico, R., and Zandipour, M. (**2000**). "A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss," J. Phonetics **28**, 233–272.

Perkell, J., Matthies, M., Lane, H., Guenther, F., Wilhelm-Tricarico, R., Wozniak, J., and Guiod, P. (**1997**). "Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models," Speech Commun. **22**, 227–250.

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., Stockmann, E., and Guenther, F. H. (**2004**). "The distinctness of speakers /s/-/ʃ/ contrast is related to their auditory discrimination and use of an articulatory saturation effect," J. Speech Lang. Hear. Res. **47**, 1259–1269.

Rabinov, C. R., Kreiman, J., Gerratt, B. R., and Bielamowicz, S. (**1995**). "Comparing reliability of perceptual ratings of roughness and acoustic measure of jitter," J. Speech Lang. Hear. Res. **38**, 26–32.

Richardson, L. M., Busby, P. A., Blamey, P. J., Dowell, R. C., and Clark, G. M. (**1993**). "The effects of auditory feedback from the Nucleus cochlear implant on the vowel formant frequencies produced by children and adults," Ear Hear. **14**, 339–349.

Svirsky, M. A., and Tobey, E. A. (**1991**). "Effect of different types of auditory stimulation on vowel formant frequencies in multichannel cochlear implant users," J. Acoust. Soc. Am. **89**, 2895–2904.

Svirsky, M. A., Lane, H., Perkell, J. S., and Wozniak, J. (**1992**). "Effects of short-term deprivation on speech production in adult cochlear implant users," J. Acoust. Soc. Am. **92**, 1284–1300.

Tartter, V. C., Chute, P. M., and Hellman, S. A. (**1989**). "The speech of a postlingually deafened teenager during the first year of use of a multichannel cochlear implant," J. Acoust. Soc. Am. **86**, 2113–2121.

Tjaden, K., and Turner, G. S. (**1997**). "Spectral properties of fricatives in Amyotrophic Lateral Sclerosis," J. Speech Lang. Hear. Res. **40**, 1358–1372.

Tobey, E. A., Gray, R., Simpson, A., Gabbert, G., Guravitch, S., and Kunkel, F. (**2000**). "Temporal characteristics of speech in young French-speaking children with cochlear implants," in *Cochlear Implants*, edited by S. B. Waltzman and N. L. Cohen (Thieme, New York), pp. 314–315.

Tye-Murray, N., Spencer, L., Bedia, E. G., and Woodworth, G. (**1996**). "Differences in children's sound production when speaking with a cochlear implant turned on and turned off," J. Speech Lang. Hear. Res. **39**, 604–610.

Uchanski, R. M., and Geers, A. E. (**2003**). "Acoustic characteristics of the speech of young cochlear implant users: A comparison with normal-hearing age-mates," Ear Hear. **24**, 90S–105S.

Xu, Y., Larson, C. R., Bauer, J. J., and Hain, T. C. (**2004**). "Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences," J. Acoust. Soc. Am. **116**, 1168–1178.

Yu, F., Revis, J., Wuyts, F. L., Zanaret, M., and Giovanni, A. (**2002**). "Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale," Folia Phoniatr Logop **54**, 271–281.

Fundamental frequency of infants' and parents' utterances in longitudinal recordings

Shigeaki Amano, Tomohiro Nakatani, and Tadahisa Kondo
NTT Communication Science Laboratories, NTT Corporation, 2-4 Hikari-dai, Seika-cho, Souraku-gun, Kyoto 6190237, Japan

(Received 27 February 2005; revised 24 September 2005; accepted 5 December 2005)

The fundamental frequencies ($F0$) of daily life utterances of Japanese infants and their parents from the infant's birth until about 5 years of age were longitudinally analyzed. The analysis revealed that an infant's $F0$ mean decreases as a function of month of age. It also showed that within- and between-utterance variability in infant $F0$ is different before and after the onset of two-word utterances, probably reflecting the difference between linguistic and nonlinguistic utterances. Parents' $F0$ mean is high in infant-directed speech (IDS) before the onset of two-word utterances, but it gradually decreases and reaches almost the same value as in adult-directed speech after the onset of two-word utterances. The between-utterance variability of parents' $F0$ in IDS is large before the onset of two-word utterances and it subsequently becomes smaller. It is suggested that these changes of parents' $F0$ are closely related to the feasibility of communication between infants and parents. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2161443]

PACS number(s): 43.70.Ep, 43.70.Gr [AL] Pages: 1636–1647

I. INTRODUCTION

A. Infant *F0*

An infant's speech development can be studied using several approaches. One involves analyzing infant utterances acoustically to reveal the developmental changes that occur with age. Among the acoustic characteristics of utterances, we focused on the fundamental frequency ($F0$). Several variables have been used in previous research to represent the $F0$ change that occurs during speech development. These variables include mean (Robb and Saxman, 1985), mode (Laufer and Horii, 1977), median (Laufer and Horii, 1977), range (Keating and Buhr, 1978), maximum (Kent and Bauer, 1985), minimum (Fairbanks, 1942), and standard deviation (Robb and Saxman, 1985). The number of percentage of occurrences has also been used to determine the developmental changes in, for example, rising, falling, and flat $F0$ patterns (Robb, Saxman, and Grant, 1989). Of these variables, we used the mean and standard deviation of $F0$ in this study, because the mean has frequently been used in previous studies and the standard deviation would be a better and more robust indicator than the range to represent $F0$ variability (cf. Masataka, 2002).

Kent (1976) reviewed the developmental change in the $F0$ mean from birth to adulthood by collecting the results of cross-sectional studies. The infant $F0$ is high at birth and gradually becomes lower as the infant grows up. The rate at which $F0$ decreases is higher before 3 years of age than after it. A difference appears in the $F0$ of boys and girls after the adolescent voice change of boys at 11 or 12 years of age. As Kent (1976) admitted, it was a tentative view of $F0$ changes with the limited data available at that time. The actual $F0$ mean and its developmental change might differ from that presented in his review.

More recently, some studies have reported the precise $F0$ mean and its developmental change. For example, Bennet (1983) reported the developmental change in $F0$ in the 7–11 age range based on 3 years of longitudinal recordings. She reported that the $F0$ mean decreases about 12 Hz per 12 months. Hollien, Green, and Massey (1994) studied the $F0$ mean of boys in their early teens around the adolescent period with 5 years of longitudinal recordings. According to their study, the $F0$ mean for 34 boys was 233.1 Hz before adolescence, 173.8 Hz during adolescence, and 121.6 Hz after adolescence. This drop in the $F0$ mean was also shown by the Lee, Potamianos, and Narayanan (1999) cross-sectional study for 436 children in the 5–17 age range. They reported that pubertal $F0$ change starts about age 12 and ends about age 15 for male children. They also reported that $F0$ significantly decreases in the 7–12 age range for male and female children but not after age of 12 for female children.

The $F0$ mean was also reported for younger ages. For example, Robb, Saxman, and Grant (1989) longitudinally recorded the utterances of seven infants in the 8–26 month age range on 12 occasions over 13 months. They investigated the $F0$ of monosyllabic and bisyllabic utterances but they found no significant difference between months for either type of utterance. The $F0$ mean was 396 Hz for monosyllables and 399 Hz for bisyllables. McRoberts and Best (1997) studied the $F0$ mean of one infant of 3–17 months under three conditions, infant alone, infant with mother, and infant with father. The $F0$ was not dependent on the conditions but changed significantly with age. That is, there was a general tendency for the $F0$ mean to decrease month by month. It was 407 Hz at 7 months, 330 Hz at 15 months, and 326 Hz at 17 months.

However, longitudinal studies in infant and early childhood were conducted over much shorter periods than those in late childhood. The longest study undertaken with infants and in early childhood is the Robb, Saxman, and Grant (1989) study consisting of observations over 1.5 years. As

 0001-4966/2006/119(3)/1636/12/$22.50

regards late childhood, a 3-year research project was undertaken by Bennet (1983) and a 5-year study was undertaken by Hollien, Green, and Massey (1994). As a consequence, few longitudinal data have been provided for *F*0 from birth to about 5 or 6 years old. If data for this range are provided and combined with results such as those of the Lee, Potamianos, and Narayanan (1999) study in 5–17 age range, we can obtain a consecutive view of the developmental change in *F*0 from birth to young adulthood.

As regards *F*0 variability, previous studies used two kinds of standard deviation (SD). One is the within-utterance SD of *F*0. This is an average of the SD of *F*0 in each utterance. This variable represents *F*0 variability within one utterance. The other is the between-utterance SD of *F*0. This is the SD of the *F*0 mean. This variable represents the variability of the *F*0 mean among utterances.

Prescott (1975) used these two kinds of SD to analyze the *F*0 of infants. He found that the within-utterance SD of *F*0 observed longitudinally increases from 28 Hz at 0 months to 53 Hz at 6–9 months for the same four infants. Moreover, cross-sectional observations revealed that the within-utterance SD of *F*0 also increases from 30 Hz at 1 month to 53 Hz at 6–8 months with 10 infants for each month. On the other hand, he found no significant increase or decrease in the between-utterance SD of *F*0 in either case. Shepard and Lane (1968) reported that the within- and between-utterance variability of *F*0 were both unchanged in a longitudinal observation of two infants from 0–4.5 months. Robb, Saxman, and Grant (1989) showed that the between-utterance SD of *F*0 did not change during an approximately 1-year longitudinal observation of infants in the 8–26 month age range. Robb and Saxman (1985) conducted cross-sectional research with 14 infants. They reported that the between-utterance SD in *F*0 is much higher at 11–16 months than at 17–25 months.

These studies offer some information about the developmental change in *F*0 variability. However, they do not provide sufficient data to depict the developmental change in *F*0 variability from birth to early childhood because the observation period is too short. In addition, they provided contradictory results in certain cases. For example, Prescott (1975) reported an increase in the within-utterance *F*0 variability but Shepard and Lane (1968) did not. These problems should be resolved by investigating *F*0 data in early childhood.

B. Parents *F*0

We also investigate another developmental aspect of *F*0, namely "infant-directed speech (IDS)," which is spoken to an infant by a caretaker such as a parent. IDS is sometimes called motherese (Shute and Wheldall, 1989), maternal speech (Penman *et al.*, 1983), infant directed talk (Werker and McLeod, 1989), or baby talk (Singh, Morgan, and Best, 2002). IDS is different from "adult-directed speech (ADS)" in some respects. For example, it has a higher *F*0 (McRoberts and Best, 1997), a wider *F*0 range (Fernald *et al.*, 1989), a shorter utterance duration (Fernald and Simon, 1984), longer pauses (Grieser and Kuhl, 1988), and an exaggerated formant frequency (Kuhl *et al.*, 1997). IDS with these characteristics attracts an infant more than ADS (Cooper *et al.*, 1997; Werker and McLeod, 1989). And this preference probably helps an infant to acquire its native language (Hirsh-Pasek *et al.*, 1987).

Within the characteristics of IDS, *F*0 seems to be an important factor in terms of an infant's preference. The work of Spence and Freeman (1996) supports this idea. They compared an infant's preference as regards low-pass filtered IDS and whispered IDS. The low-pass filtered IDS contains *F*0 information but the whispered IDS does not. Infants preferred the low-pass filtered IDS to the whispered IDS, suggesting that *F*0 is an important factor in IDS. Additional support for this idea was obtained by Fernald and Kuhl (1987). They reported that infants prefer synthesized IDS with *F*0 modulation to that with amplitude modulation. This suggests that *F*0 in IDS is an outstanding feature for infants.

Many of the characteristics of IDS are said to be found in almost every language, and it is suggested that these characteristics might be universal (e.g., Grieser and Kuhl, 1988). But there are some exceptions. For instance, Ratner and Pye (1984) reported that a higher *F*0 in IDS is not observed in Quiche Mayan, which is spoken in the western highlands of Guatemala. In addition, the extent of *F*0 change seems to differ across languages. For example, mothers of American-English native speakers tend to use a much more exaggerated *F*0 in IDS than mothers who speak other languages (e.g., Fernald *et al.*, 1989; Papousek, Papousek, and Symmes, 1991). Similarly, mothers of Australian-English native speakers tend to use a more exaggerated *F*0 in IDS than mothers of Thai native speakers (Kitamura *et al.*, 2002). Shute and Wheldall (1989) reported that mothers of British-English native speakers use a high *F*0 in IDS, but that the increase in *F*0 is less than for mothers of American-English native speakers.

As for Japanese, IDS seems to have the same characteristics as other languages. For example, Niwano and Sugai (2003) reported that mother and father use a higher *F*0 in terms of mean, maximum, minimum, and range in IDS to infants at 3, 5, and 7 months of age. Masataka (1992) showed that there is a higher mean and a wider range of *F*0 in IDS than in ADS especially when infants do not respond to the IDS. Masataka (2002) reported that the *F*0 of IDS to an 11-month-old infant is higher than that of ADS. He also found that the *F*0 increase in IDS is larger in a female with children (i.e., mother) than a female without children (i.e., nonmother).

Several factors affect the *F*0 in IDS. One relates to people's experience with their own babies as shown by Masataka (2002). Other factors are, for example, an infant's facial expressions (Reissland, Shepherd, and Cowie, 2002) and the cradling side (Reissland, 2000). Although there may be many other factors including language difference, previous studies almost always provide evidence that *F*0 of IDS is high.

However, it is not clear how *F*0 in IDS changes with infant age. There have only been a few studies on *F*0 changes in IDS with age (e.g., Remick, 1976; Warren-Leubecker and Bohannon, 1984). For example, Remick (1976) studied mothers' *F*0 in IDS to their infants whose ages ranged between 16 and 30 months with an approxi-

mately 2-month interval. Although she showed the *F*0 change in IDS as a function of an infant's age, the observed *F*0 change might be an artifact of individual differences, because she used a cross-sectional method. Recently, Kitamura *et al.* (2002) observed the *F*0 change in IDS with a longitudinal method for infants from 0 to 12 months old. They showed that the *F*0 in IDS is higher than in ADS and that it changes with infant age. However, their observation period was less than 1 year. It is still unknown how *F*0 in IDS changes over longer periods. A greater change may be found if the *F*0 in IDS is observed with infants for more than 1 year. In particular, long term observation would enable us to discover when the special characteristics of *F*0 in IDS disappear and become the same as in ADS. Little attention has been paid to this point as regards Japanese or other languages. It was simply assumed that IDS has some special characteristics in relation to a prevabal infant (e.g., Sachs, 1977).

It is likely that *F*0 changes in IDS are closely related to the speech development of infants, and that a large *F*0 change occurs in IDS when the infant starts to use "language." We assume that this change would occur at the onset of two-word utterances, because then parents can communicate more easily, and probably more normally, with their infant using "language." This idea is supported by the fact that Japanese mother-child conversation style becomes adult-like around the two-word utterance period (Kajikawa, Amano, and Kondo, 2004). The onset of one-word utterances is another possible change point. However, we assume that parents have a lot of difficulty communicating with their infant at the onset of one-word utterances, and would not speak to the infant in the same way as to adults. Based on this idea, we analyzed the *F*0 in IDS before and after the onset of two-word utterances. The mean and standard deviation were used for the analysis as with infant *F*0. The infant's *F*0 was also analyzed before and after the onset of two-word utterances, because this supplies additional information about the *F*0 change in infant utterances, which might relate to the *F*0 change in the IDS of parents.

In summary, we try to clarify the developmental changes that take place in the *F*0 mean and its variability in infants' and parents' utterances before and after the onset of two-word utterances. It would provide systematic information on *F*0 development with age from 0 to 5 years old, which has not been well studied in previous research.

II. *F*0 ANALYSIS

To observe the developmental change in *F*0, we conducted two *F*0 analyses of the utterances of an infant and its parents. The first was intended to show that the *F*0 of infant utterances decreases month by month. It was also intended to show the developmental change in the *F*0 variability of infant utterances in terms of between- and within-utterance SD. More specifically, we compared these values before and after the onset of two-word utterances.

The second *F*0 analysis concerned the fact that Japanese parents' *F*0 is different in IDS and ADS. We intended to show that the *F*0 mean in IDS is high before the onset of an infant's two-word utterances and it has almost the same value as ADS after it. We also intended to show developmental change in *F*0 variability in IDS with regard to between- and within-utterance SD. More specifically, we compared these values before and after the onset of two-word utterances and in ADS.

A. Infant *F*0

1. Data

An infant speech database (Amano, Kato, and Kondo, 2002) was used as the data population. The database contains speech files (16-bit quantization and a 16-kHz sampling frequency) of natural utterances of five pairs of Japanese infants and parents during daily life at home. The utterances were digitally recorded approximately monthly for about 1 hour from birth to 5 years old. Infants B, C, and D were selected from the infant speech database, because the number of registered utterances was larger for these infants than the others.

All the infants were born and raised in Tokyo or its suburbs. They were all of normal height and weight at birth, and had no symptoms of disorder with respect to speech perception or speech production during the recording period. Although Infants C and D are siblings, we regarded their data as independent because their recordings were conducted separately.

The utterances that satisfied the following conditions were extracted from the database and used for analysis:

(i) The background noise should be low so that utterances are clear.
(ii) There should be no overlap with other utterances so that *F*0 value is clear.
(iii) Utterances with certain particular characteristics must be excluded in order to focus on normal utterances. Such characteristics include hiccupping, coughing, sneezing, yawning, singing, reading aloud, and number counting.

Table I shows the months during which infant utterances were extracted. The mean value of the number of extracted utterances per month was 432 (SD=312) for Infant B, 406 (SD=173) for Infant C, and 486 (SD=243) for Infant D.

2. Procedure

A voiced/unvoiced part was specified in the extracted utterances by using the dominance spectrum method, called "DominAnce Spectrum based Harmonics extraction (DASH)" (Nakatani, Irino, and Zolfaghari, 2003; Nakatani and Irino, 2004). *F*0 was estimated for the voiced part by using the "Ripple Enhanced Power Spectrum (REPS)" method (Nakatani and Irino, 2002, 2004). A 42 ms Hanning window with 1 ms window shift was used for both methods.

Although these methods perform very well in terms of voiced/unvoiced (V/UV) classification and *F*0 estimation, there are occasional errors. To reduce the errors as much as possible, trained operators checked and corrected the *F*0 values and V/UV labels by investigating the superimposed *F*0 and V/UV information segment by segment on a spectrogram.

TABLE I. Months containing infant's utterances in infant speech database.

Infant ID	Month	Number of months
B	0, 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 33, 34, 35, 36, 37, 39, 40, 42, 43, 44, 45, 48, 51, 52, 53	45
C	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 47, 48, 49, 50, 53, 55, 56, 59, 60	52
D	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 43, 44, 45, 46, 47, 48, 49, 51, 52, 55, 56, 59	53

To exclude unusual $F0$ values, we used those between 20 and 1500 Hz. Although these lower and higher limits were decided empirically, they are appropriate for covering the $F0$ range of an infant utterance. The suitability of the limits is supported by, for example, Robb and Saxman (1985). They reported that the $F0$ mean was 357 Hz, the SD was 105 Hz, and the range of the mean was 164–1366 Hz for the utterances of 14 infants between the ages of 11 and 25 months.

$F0$ mean, SD (within) and SD (between) were calculated from the $F0$ value for the voiced segment. $F0$ mean is the average of the mean $F0$ value of each utterance. SD (within) is the average of standard deviation of the $F0$ value of each utterance. SD (between) is the standard deviation of the $F0$ mean value among utterances. SD (within) represents the $F0$ variability within one utterance, whereas SD (between) represents the variability of the $F0$ mean value among utterances.

To examine the developmental change in infant' $F0$, regression analysis was conducted for the $F0$ mean as a dependent variable (y) with infant's age in months as an independent variable (x). The regression method was the best linear unbiased estimation.

We undertook analyses of variance for the two SDs to compare their values before and after the onset of two-word utterances. The onset of two-word utterances was decided by two Japanese adults by listening to recordings of infants' utterances. The definition of a two-word utterance in this study is an utterance that contains two content words whose pronunciation is clear enough to be recognized by both Japanese adults.

3. Results and discussion

$F0$ mean in infant utterances: Figure 1 shows the $F0$ mean of infant utterances as a function of month. The regressions of the $F0$ mean were $y=368-0.65x$ [$r^2=0.109$, $F(1,43)=5.27$, $p<0.05$] for Infant B, $y=373-1.53x$ [$r^2=0.687$, $F(1,50)=109.67$, $p<0.001$] for Infant C, and $y=384-1.11x$ [$r^2=0.555$, $F(1,51)=63.59$, $p<0.001$] for Infant D. The coefficient of x of the regression was significant for all infants. [$t(43)=-2.30$, $p<0.05$, two-tailed, for Infant B; $t(50)=-10.47$, $p<0.001$, two-tailed, for Infant C; $t(51)=-7.97$, $p<0.001$, two-tailed, for Infant D]. When the data was pooled for all the infants, the regression of the $F0$ mean was $y=385-1.33x$ [$r^2=0.400$; $F(1,148)=98.75$, $p<0.001$]. The coefficient of x of the regression was significant [$t(148)=-9.94$, $p<0.001$, two-tailed].

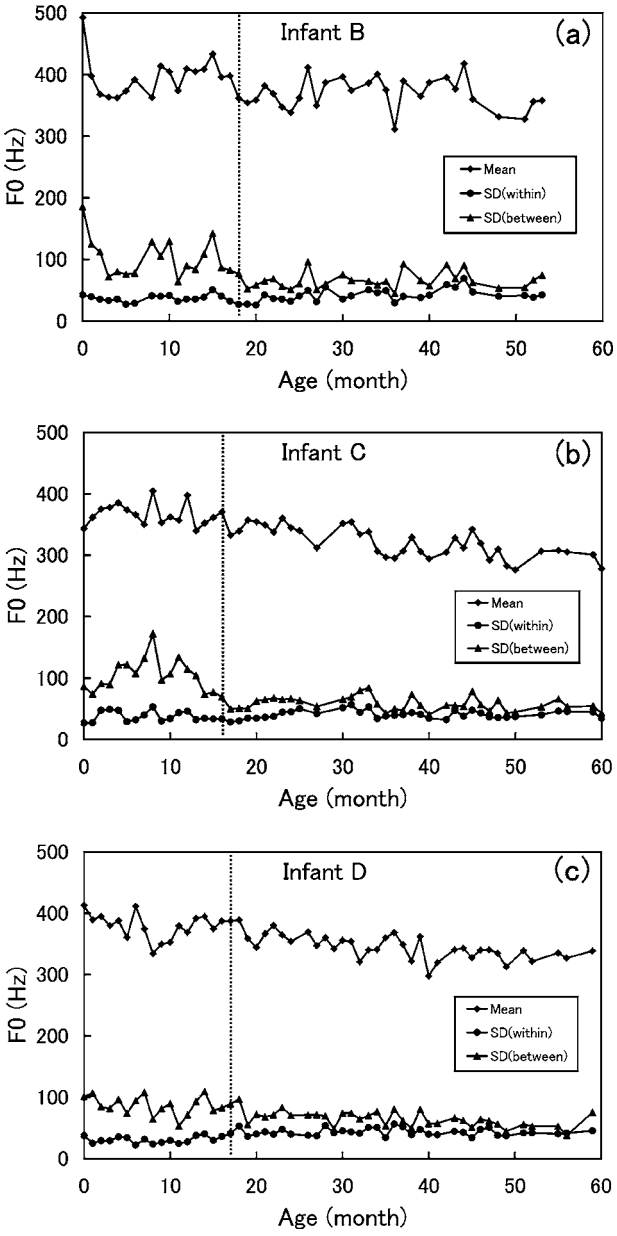


FIG. 1. $F0$ mean and SD of infant utterances as a function of the infant's age in months. The vertical dotted line represents the onset of two-word utterances for each infant.

These negative and significant coefficients indicate that $F0$ has a tendency to decrease month by month both for an individual infant and infants as a whole. This tendency is consistent with the results of previous studies (e.g., Bennet, 1983; Eguchi and Hirsh, 1969; Kent and Murray, 1982; McRoberts and Best, 1997). In addition, the rate of $F0$ de-

TABLE II. SD (between) and SD (within) of *F*0 of infant utterances before/after onset of two-word utterances.

	SD (between) (Hz)				SD (within) (Hz)			
	Infant ID				Infant ID			
	B	C	D	Mean	B	C	D	Mean
Before	103.9	106.6	86.5	98.6	37.2	38.1	30.7	35.2
After	66.1	58.3	65.2	63.0	41.9	40.6	43.3	42.0

crease is similar to that reported in some previous studies. From the regression of the pooled data in our study, we obtain an *F*0 decrease rate of 16 Hz per 12 months for infants of 0–5 years old. This *F*0 decrease rate is close to the result reported by Bennet (1983). She showed that the *F*0 decrease rate is about 12 Hz per 12 months for children in the 7–11 year age range.

However, some previous studies have reported different *F*0 decrease rates. For example, when we calculate the *F*0 decrease rate from the results of Eguchi and Hirsh (1969), it is 8.9 Hz per 12 months for 3–6 year-old children. This is much smaller than our result. In addition, their data indicates that the decrease rate is almost zero at age 7–9 years. Although this difference might be an artifact of the cross-sectional method that Eguchi and Hirsh (1969) used, it is possible that the *F*0 decrease rate changes with age in childhood. At least for boys, the *F*0 value drops very sharply during the period of adolescent voice change between about 12 and 15 years of age (Hollien, Green, and Massey, 1994; Lee, Potamianos, and Narayanan, 1999). However, we currently have insufficient evidence about the change in the *F*0 decrease rate prior to the period of adolescent voice change. We may only say from our data that *F*0 has a general tendency to decrease month by month and that the decrease rate is constant from 0–5 years of age.

Assuming that *F*0 is about 300–400 Hz for infants, the *F*0 decrease rate per 12 months is about 3.0%–4.0% in Bennet's (1983) study, about 2.2%–3.0% in the Eguchi and Hirsh (1969) study, and about 1.9%–6.1% in the current study. Because these percentages are very small, it would be difficult to detect the tendency of the *F*0 decrease when speech samples are collected over a short period. This might be one reason why some previous studies failed to find any significant tendency in the *F*0 decrease. For example, in Prescott's (1975) study, no consistent increase or decrease tendency was observed for *F*0 mean between 0 months and 6–9 months either in a longitudinal same-infant group or in a cross-sectional different-infant group. Robb, Saxman, and Grant (1989) conducted longitudinal recordings 12 times over 13 months with seven children 8–26 months of age, but they did not find any significant *F*0 difference among months. Robb and Saxman (1985) also failed to find any significant *F*0 decrease when using the cross-sectional method with 14 infants 11–25 months of age. The observation period was less than 1.5 years in all these studies. It is suggested that 2 or more years of observation would be necessary to obtain a significant tendency for the *F*0 decrease.

As regards *F*0 changes during the first 12 months, Shepard and Lane (1968) reported that the *F*0 mean decreased between 0 and about 1 month and then increased and became stable at 2–4.5 months. On the other hand, Fairbanks (1942) showed that the *F*0 mean increases at 0–4 months and then becomes almost constant at 5–9 months. Kent and Murray (1982) reported that the *F*0 mean of infants is 445 Hz at 3 months, 450 Hz at 6 months, and 415 Hz at 9 months. These short-period *F*0 changes are not confirmed by our *F*0 analysis in this study, because the *F*0 mean fluctuates from month to month and its pattern differs among infants (Fig. 1). For example, the *F*0 mean decreases at 0–4 months for Infants B and D, but it increases for Infant C. Such individual differences were also observed by Laufer and Horii (1977) in longitudinal *F*0 observations of four infants over the first 24 weeks. As they pointed out, individual *F*0 fluctuation depends on many factors including infant's activity level, utterance skill, and physical and social environment. These factors may differ among studies. Therefore, it would be difficult to verify the true *F*0 change over a short period. To reveal the *F*0 change over a short period, further research is necessary with precisely controlled conditions as regards the infant and his/her environment.

*F*0 *variability in infant utterances*: Figure 1 shows SD (between) and SD (within) of *F*0 in infant utterances as a function of month. Table II shows the averaged SD (between) and SD (within) before and after the onset of two-word utterances. We performed an analysis of variance with one factor of period with two levels (before vs after the onset of two-word utterances). An infant factor was included in a residual in the analysis. The factor of period was significant for SD (between) $[F(1,146)=122.3,\ p<0.001]$. This indicates that SD (between) is larger before the onset of two-word utterances than after it. Another analysis of variance showed that the factor of period was also significant for SD (within) $[F(1,146)=26.8,\ p<0.001]$. This indicates that SD (within) is smaller before the onset of two-word utterances than after it.

These results mean that infant *F*0's are very different from each other between utterances before the onset of two-word utterances, but this difference is reduced after the onset of two-word utterances. On the other hand, infant *F*0 does not change very much within one utterance before the onset of two-word utterances, but it becomes more variable within one utterance after the onset of two-word utterances.

The developmental change in *F*0 variability probably reflects an infant's ability to speak a language. Infants do not speak a language when they are very young. During that

period, it is highly probable that infants do not have the ability to control *F*0 precisely. This inability to control *F*0 would mean that infants could not start an utterance with a particular *F*0. At the same time, they could not quickly increase or decrease the *F*0 within one utterance. These characteristics are reflected in the large between-utterance variability and the small within-utterance variability before the onset of two-word utterances. It might be said that infant *F*0 is physically driven with some randomness, rather than intentionally, before the onset of two-word utterances.

As infants grow up, they acquire the ability to control *F*0 precisely. This acquisition of *F*0 control is critical with respect to spoken language, because the *F*0 pattern conveys a message in communication. For example, interrogative and affirmative information is expressed by the *F*0 pattern. In addition, the *F*0 pattern covers part of the lexical information in Japanese (Kubozono, 1993). With precise *F*0 control, infants can intentionally start their utterances with a fairly constant *F*0. And they can quickly increase or decrease the *F*0 within one utterance to create a complex *F*0 pattern when communicating with other human beings, mainly with their parents. These characteristics are reflected in the decrease in between-utterance variability and the increase in within-utterance variability after the onset of two-word utterances.

Robb and Saxman (1985) obtained very similar results to ours regarding SD (between), although they used the cross-sectional method with 14 infants. They showed that SD (between) is large (63–238 Hz) at 11–16 months and it becomes small (53–84 Hz) at 17–25 months. Because the onset of two-word utterances corresponds to 17–19 months in our study, their results agree with ours in that SD (between) is large before the onset of two-word utterances and subsequently becomes small. In addition, our SD (between) values before and after the onset of two-word utterances are in a similar range to theirs. These correspondences support our claim that the between-utterance *F*0 variability is large before the onset of two-word utterances and subsequently small.

However, some studies appear to disagree with our result. For example, Robb, Saxman, and Grant (1989) showed that SD (between) does not change between 8 and 26 months of age. This means that the between-utterance *F*0 variability is constant before and after the onset of two-word utterances (about 17–19 months). However, their study is methodologically different from ours. They observed only one- and two-syllable utterances, whereas we observed almost all types of utterance including nonlinguistic utterances. Before the onset of two-word utterances, there are other types of utterance in addition to one- and two-syllable utterances, because nonlinguistic utterances cannot always be categorized as "*n*-syllable" utterances. Such uncategorized utterances are probably dominant and they have a greater variety of *F*0 values than one- and two-syllable utterances. If the uncategorized utterances are excluded, the between-utterance *F*0 variability would be greatly reduced. In addition, after the onset of two-word utterances, the uncategorized utterances would become less common, and linguistic utterances such as one- and two-syllable utterances would become dominant. Because the one- and two-syllable utterances would have less *F*0 variability than the uncategorized utterances, the between-utterance *F*0 variability is small after the onset of two-word utterances. As a consequence, the between-utterance *F*0 variability would be small both before and after the onset of two-word utterances, and no difference would be observed between them in the Robb, Saxman, and Grant (1989) study. Therefore, the Robb, Saxman, and Grant (1989) results would not necessarily contradict our results. Rather, their different results are probably caused by the limits of their utterance observation.

As for the within-utterance *F*0 variability, Prescott (1975) reported that SD (within) is 28 Hz at 0 months and 53 Hz at 6–9 months. Our SD (within) before the onset of two-word utterances (35.2 Hz) falls in this range. Laufer and Horii (1976) reported that SD (within) is 27.7–38.8 Hz at 0–5.5 months. Our SD (within) again falls in their range. These coincidences with previous studies support the validity of our results for within-utterance *F*0 variability before the onset of two-word utterances. However, because their observation periods were much shorter than ours, the support is limited. No other studies have investigated the within-utterance *F*0 variability after 10 months of age. Specifically, no studies have investigated the difference in the within-utterance *F*0 variability before and after the onset of two-word utterances. Further study is needed to check the validity of our results for within-utterance *F*0 variability.

One of the claims of Prescott (1975) is that SD (within) increases from 0 to 9 months. However, when we conducted regression analysis with our data between 0 and 9 months with SD (within) as a dependent variable y and month as a independent variable x, the regression coefficient of x was not significant either for an individual infant or for infants as a whole. Therefore, Prescott's (1975) claim as regards the increase in within-utterance *F*0 variability is not confirmed by the current study. Nor is it in agreement with the results of Laufer and Horii (1976) or Shepard and Lane (1968). One of the reasons for this disagreement would be that Prescott (1975) observed only an infant's cry, whereas the current study, Laufer and Horii (1976), and Shepard and Lane (1968) observed an infant's utterances including cry and noncry. It is possible that the within-utterance *F*0 variability of cry exhibits a different developmental change from that of noncry. However, the observed increase might be an artifact of fluctuation because Prescott (1975) used only two sample points (0–1 months and 6–9 months). It would be better to use more sample points along with infant age to observe any developmental change.

B. Parents *F*0

1. Data

The parents of Infants B, C, and D were selected for analysis. Infants C and D are siblings and this might cause some artifacts to appear in the data. However, because the recordings were conducted separately for Infants C and D, such artifacts would be very small. Therefore we treat them as having different parents allowing for the possibility of artifacts.

The months containing parents' utterances in the data-

TABLE III. Months containing parents's utterances in infant speech database.

Infant ID	Parent	Month	Number of months
B	Father	0, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 15, 16, 19, 20, 21, 22, 23, 24, 25, 27, 28, 30, 31, 33, 34, 37, 39, 40, 42, 43, 44, 48, 51, 53	36
	Mother	0, 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 33, 34, 35, 36, 37, 39, 40, 44, 45, 48, 51, 52, 53	43
C	Father	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 47, 48, 49, 50, 53, 55, 56, 59, 60	52
	Mother	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 45	38
D	Father	0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 43, 44, 45, 46, 47, 48, 49, 51, 52, 55, 56 59	52
	Mother	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 27, 28, 29, 31, 32, 33, 34, 39, 44, 48, 52, 59,	35

base are shown in Table III. The parents' utterances were extracted from the database under the same conditions as those used for the infants. Their IDS and ADS were separately extracted from the database. The mean number of extracted utterances is shown in Table IV.

2. Procedure

*F*0 values were obtained by the same procedure as for the infant utterances except that the upper *F*0 limit was 250 Hz for a father and 500 Hz for a mother. These limits were empirically set to exclude unusual *F*0 values. But they are probably sufficient to cover the *F*0 range of parents' utterances, because Terasawa, Kakita, and Hirano (1984) reported that the *F*0 mean of Japanese males is 122 Hz and the standard deviation is 19.9 Hz, whereas the *F*0 mean of Japanese females is 241 Hz and the standard deviation is 20.0 Hz. Moreover, *F*0 is about 110–140 Hz for male adults

TABLE IV. Mean and SD (in parentheses) of the number of parents' utterances, and the number of months used for analysis before/after onset of two-word utterances.

			Mean of number of utterances per month			Number of months			
			Infant ID			Infant ID			
			B	C	D	B	C	D	Mean
Father	IDS	Before	69 (66)	85 (67)	200 (111)	14	16	16	15.3
		After	147 (127)	171 (161)	348 (173)	22	36	36	31.3
	ADS		11 (13)	6 (5)	3 (2)	29	26	14	23.0
Mother	IDS	Before	125 (101)	257 (186)	161 (94)	17	16	16	16.3
		After	72 (72)	367 (202)	116 (94)	26	22	19	22.3
	ADS		10 (13)	5 (5)	3 (3)	31	27	7	21.7

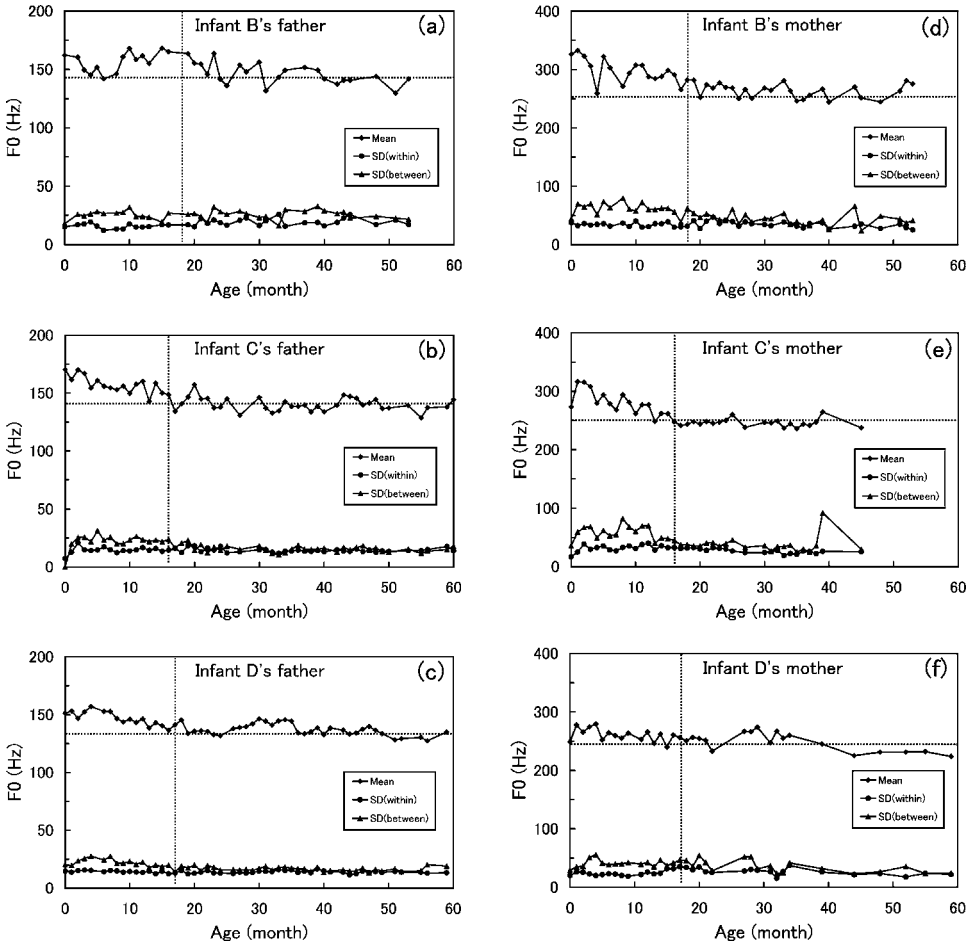


FIG. 2. *F*0 mean and SD of infant-directed speech by parents. The vertical dotted line represents the onset of two-word utterances for each infant. The horizontal dotted line represents *F*0 mean in the ADS of each parent.

and 190–230 Hz for female adults according to the summary of previous *F*0 studies provided by Hollien, Hollien, and de Jong (1997). The mean and SD of *F*0 in IDS were calculated for each month. On the other hand, the mean and SD of *F*0 in ADS were calculated for all the months. This is because ADS characteristics most probably do not change with an infant's age, and because the number of ADS utterances (Table IV) is too small to provide reliable mean and SD values for each month.

3. Results and discussion

*F*0 *mean in parent utterances:* Figure 2 shows the *F*0 mean of infant-directed speech uttered by the parents of Infants B, C, and D as a function of month. The *F*0 mean in IDS before the onset of two-word utterances tends to be higher than after it. Table V shows the *F*0 mean in IDS before and after the onset of two-word utterances and in ADS. As shown in Table V, the father's *F*0 mean before the onset of two-word utterances tends to be 10–20 Hz higher than that in ADS. Similarly, the mothers' *F*0 mean before the onset of two-word utterances tends to be 10–40 Hz higher than that in ADS. However, there seems to be no difference in *F*0 between ADS and IDS after the onset of two-word utterances.

To confirm these tendencies, we conducted an analysis of variance of *F*0 with one factor with three levels (IDS before the onset of two-word utterances, IDS after the onset of two-word utterances, and ADS). The father and mother were analyzed separately because the gender difference is obvious in an adult's *F*0 (e.g., Hollien, Hollien, and de Jong, 1997). A factor of the individual parent was included in a residual.

For the father, there was a significant differences in the *F*0 among the three levels $[F(2,204)=28.19,\ p<0.001]$. An HSD test revealed that IDS before the onset of two-word utterances is significantly different from ADS ($p<0.05$) and from IDS after the onset of two-word utterances ($p<0.05$). No significant difference was found between IDS after the onset of two-word utterances and ADS. Similar results were obtained for the mother. There was a significant difference among these three levels $[F(2,176)=38.76,\ p<0.001]$. The HSD test revealed that IDS before the onset of two-word utterances is significantly different from ADS ($p<0.05$) and from IDS after the onset of two-word utterances ($p<0.05$). No significant difference was found between IDS after the onset of two-word utterances and ADS.

These results indicate three points. The first is that, for both father and mother, *F*0 mean is higher in IDS before the onset of two-word utterances than ADS. That is, Japanese parents use a higher *F*0 when speaking to their infant than to other adults. This point agrees with previous research for Japanese (e.g., Fernald *et al.*, 1989; Masataka, 1992, 2002;

TABLE V. *F*0 mean before and after onset of two-word utterances in IDS and in ADS.

			*F*0 mean (Hz)			
			Infant ID			
			B	C	D	Mean
Father	IDS	Before	157	158	147	154
		After	147	140	137	141
	ADS		143	141	133	140
Mother	IDS	Before	298	281	261	280
		After	264	246	249	254
	ADS		254	250	246	251

Niwano and Sugai, 2003) and with other languages (e.g., Fernald and Simon, 1984; Grieser and Kuhl, 1988; Jacobson *et al.*, 1983; McRoberts and Best, 1997; Stern *et al.*, 1983).

The second point revealed by the current results is that the *F*0 mean in IDS is higher before the onset of two-word utterances than after it. That is, parents speak to their infants with a high *F*0 before the onset of two-word utterances, but not subsequently. Few studies have focused on this point and so little data is available. One exception is the work of Jacobson *et al.*, (1983). According to their data, there is no difference in the *F*0 mean of IDS for infants at 4–8 months (corresponding to the period before the onset of two-word utterances) and at 22–30 months (corresponding to the period after the onset of two-word utterances). The *F*0 mean is high in both cases, and this is not in agreement with our results.

The reason for this disagreement might be that our parents talked to their infant in a very natural situation in their home, whereas the participants in the previous study talked to their infant in an experimental room, and the conversation included the sentences required by the experimenter. This would constitute a somewhat unnatural situation. Their participants probably tried to attract infant's attention. This might have raised the participants' *F*0 when they spoke to infants not only of 4–8 months but also of 22–30 months. This possibility is partly supported by Shute and Wheldall (1989). They showed that the *F*0 mean in IDS is higher in free talk than when sentences are read. Because the intention to attract infant's attention would be greater in free talk than when reading sentences, their results suggest a relationship between intention and *F*0 mean. This possibility is also supported by Katz, Cohn, and Moore (1996). They indicated that the *F*0 mean changes according to the pragmatic category of IDS such as attention, approval, and comfort.

The third point is that the *F*0 mean in IDS reaches almost the same value as in ADS after the onset of two-word utterances. That is, parents speak to their infants with the almost same *F*0 value as when speaking to adults after infants start to produce two-word utterances.

This third point probably results from the fact that parents can normally communicate with their infants in the same way as with adults when the infants start to speak language. In other words, the development of an infant's communication ability stops parents using a high *F*0 in IDS. If we use a high *F*0 when we speak to adults, it conveys a special meaning such as anger or surprise. Using a high *F*0 is normally inappropriate when we communicate with adults. When we want to communicate normally with an infant who can speak, a high *F*0 would also be inappropriate. For this reason, parents stop using a high *F*0 in IDS and use the same *F*0 to an infant when they start to speak two-word utterances with which the infant can communicate with his/her parents. It is suggested that the degree of difficulty or impossibility of communication is one of the main factors inducing the high *F*0 in IDS. This suggestion is supported by the fact that the *F*0 mean is high in speech to elder people (Masataka, 2002) and pets (Burnham, Kitamura, and Vollmer-Conna, 2002) where communication is difficult or impossible.

In Fig. 2, the *F*0 mean in IDS before the onset of two-word utterances seems to decrease as a function of month. When we performed a regression analysis of our data before the onset of two-word utterances with *F*0 mean as a dependent variable y and month as an independent variable x, the regression coefficient of x was negative and significantly different from zero for four out of six parents [$t(14)=-3.04$, $p<0.01$, two-tailed, for Infant B's mother; $t(12)=-2.85$, $p<0.05$, two-tailed, for Infant C's father; $t(13)=-3.16$, $p<0.01$, two-tailed, for Infant C's mother; and $t(13)=-6.15$, $p<0.001$, two-tailed, for Infant D's father]. The regression of the four parents was $y=316-2.34x$ [$r^2=0.381$, $F(1,15)=9.24$, $p<0.01$] for Infant B's mother, $y=164-0.91x$ [$r^2=0.384$, $F(1,13)=8.10$, $p<0.05$] for Infant C's father, $y=293-2.35x$ [$r^2=0.416$, $F(1,14)=9.96$, $p<0.01$] for Infant C's mother, and $y=154-0.93x$ [$r^2=0.730$, $F(1,14)=37.9$, $p<0.001$] for Infant D's father. When regression analysis was conducted for all the fathers and mothers together, the regression coefficient of x was also negative and significantly different from zero [$t(42)=-2.15$, $p<0.05$, two-tailed, for the fathers; $t(46)=-2.27$, $p<0.05$, two-tailed, for the mothers]. The regression was $y=157-0.54x$ [$r^2=0.097$, $F(1,43)=4.63$, $p<0.05$] for the fathers; and $y=282-1.28x$ [$r^2=0.099$, $F(1,47)=5.15$, $p<0.05$] for the mothers. Therefore, we can say that the *F*0 mean of IDS gradually decreases month by month before the onset of two-word utterances. Finally it reaches the same value as ADS as described earlier in this section.

Previous studies partly support this change of *F*0 mean

TABLE VI. SD (between) and SD (within) of *F*0 before and after onset of two-word utterances in IDS and in ADS.

			SD (between) (Hz)				SD (within) (Hz)			
			Infant ID				Infant ID			
			B	C	D	Mean	B	C	D	Mean
Father	IDS	Before	25.2	22.1	22.1	23.1	15.7	14.5	14.2	14.7
		After	25.7	15.8	16.5	18.3	19.4	14.4	13.9	15.4
	ADS		22.5	11.4	7.49	15.3	16.7	15.4	12.2	15.3
Mother	IDS	Before	61.7	58.6	40.7	53.8	34.2	31.8	23.7	30.0
		After	44.0	37.8	35.8	39.6	34.5	27.2	27.1	30.0
	ADS		29.1	23.9	11.3	25.0	37.1	28.9	17.9	31.6

in IDS before the onset of two-word utterances. For instance, in the McRoberts and Best (1997) longitudinal observation of IDS at 3, 7, 10, 15, and 17 months of age, the *F*0 mean of the mother tends to decrease linearly as the infant's age increases. This result coincides with our observations. However, the *F*0 mean of the father does not show a clear tendency to decrease. That is, the father's IDS changes nonlinearly as the infant's age increases.

On the other hand, some studies (e.g., Reissland, 1998; Niwano and Sugai, 2003) did not show this tendency for the *F*0 mean to decrease. In particular, Kitamura *et al.* (2002) reported a linear increase accompanied by an inverted U shape trend in the *F*0 mean of IDS. However, their trend analysis might be less reliable than ours, because their observation period (0–12 months) is shorter than ours (0–19 months) and they used fewer observation points (5 months) than us (14–17 months, see Table IV). Moreover, because the high *F*0 in IDS must fall to a normal level at some infant age, it is unreasonable to expect the trend to constitute a global increase. However, a local increase would be possible. The Kitamura *et al.* (2002) result can be interpreted as a combination of a local increase and a global decrease, because such a combination can realize an inverted U shape trend, as with their results.

We might have missed finding such a local trend in our analysis, because the *F*0 mean fluctuates as seen in Fig. 2. The problem is to determine whether such a local trend really exists, and if so, at which age it exists. These remain open questions. However, currently we can say that the *F*0 mean exhibits a global tendency to decrease as the infant's age increases in months before the onset of two-word utterances.

*F*0 *variability in parent utterances:* Figure 2 shows SD (between) and SD (within) of infant-directed speech by the parents of Infants B, C, and D as a function of month. Table VI shows SD (between) and SD (within) of the *F*0 in IDS before and after the onset of two-word utterances and in ADS. In Table VI, SD (between) in IDS before the onset of two-word utterances tends to be larger than that in IDS after the onset of two-word utterances. And SD (between) in IDS after the onset of two-word utterances tends to be larger than that in ADS. These tendencies were observed for both father and mother. On the other hand, SD (within) seems to have no obvious tendency.

To test these tendencies, we conducted a one-factor analysis of variance for SD with three levels (IDS before the onset of two-word utterances, IDS after the onset of two-word utterances, and ADS). The father and mother were analyzed separately as in the analysis of the *F*0 mean in the preceding section. A factor of individual parent was included in a residual.

For the father, the main factor was significant for SD (between) $[F(2,206)=14.46,\ p<0.001]$. An HSD test revealed that every pair of levels was significantly different for SD (between) ($p<0.05$). However, the main factor was not significant for SD (within). Similar results were obtained for the mother. The main factor was significant for SD (between) $[F(2,178)=42.63,\ p<0.001]$. An HSD test revealed that every pair of levels was significantly different for SD (between) ($p<0.05$). However, the main factor was not significant for SD (within).

The results reveal four points about *F*0 variability in parents' utterances. The first is that the between-utterance *F*0 variability is larger in IDS than in ADS both before and after the onset of two-word utterances. The second is that the between-utterance *F*0 variability in IDS is larger before the onset of two-word utterances than after it. The third is that the within-utterance *F*0 variability in IDS is the same as in ADS both before and after the onset of two-word utterances. The fourth is that the within-utterance *F*0 variability in IDS shows no difference between before and after the onset of two-word utterances.

There have been very few previous studies of *F*0 variability in IDS, and so little comparable data is available. One exception is the data provided by Jacobson *et al.* (1983). They showed that SD (within) in IDS for infants at 4–8 months is almost the same as that for infants at 22–30 months. Their results are consistent with the fourth point, because 4–8 months and 22–30 months, respectively, correspond to before and after the onset of two-word utterances. The other three points have not been previously studied and are presented here for the first time.

The four points mentioned above indicate that the extent of *F*0 in IDS and ADS is almost the same within one utterance. Moreover, the *F*0 in IDS and ADS does not change very much within one utterance. However, the *F*0 varies greatly between utterances. This is because the whole *F*0 of

an utterance is shifted upwards or downwards. This between-utterance $F0$ variability in IDS decreases after the onset of two-word utterances. One of the causes of this decrease is the difficulty or impossibility of communication as mentioned for the $F0$ mean in the preceding section. However, unlike the $F0$ mean, the between-utterance $F0$ variability in IDS does not reach the same value as that in ADS. The between-utterance $F0$ variation in IDS is larger than that in ADS even after the onset of two-word utterances. This means that parents speak to infants with a variety of $F0$ values between utterances even after the onset of two-word utterances at least up to 5 years old. This characteristic of $F0$ variability might be observed only for Japanese. Further studies in other languages are necessary in this regard.

III. GENERAL DISCUSSION

The $F0$ characteristics of Japanese infants and parents were investigated using an infant speech database. The $F0$ characteristics were analyzed for a large number of utterances over a very long observation period by monitoring the same infants and parents. This study not only confirmed some of the results of previous research on the $F0$ characteristics of infants and parents but also revealed new facts about the developmental change in $F0$ with months of age and its relationship to the onset of two-word utterances in terms of mean and between- and within-utterance variability.

As regards an infant's $F0$, the analysis showed that (1) the mean has a tendency to decrease month by month, (2) the between-utterance variability before the onset of two-word utterances is larger than that after the onset of two-word utterances, and (3) the within-utterance variability before the onset of two-word utterances is smaller than that after the onset of two-word utterances. This difference in variability probably reflects the characteristics of linguistic and nonlinguistic utterances of infants.

With respect to the parents' $F0$, the analysis revealed that (1) the mean is higher in IDS before the onset of two-word utterances than in ADS, (2) the mean in IDS is higher before the onset of two-word utterances than after it, (3) the mean in IDS becomes almost the same as in ADS after the onset of two-word utterances, and (4) the mean in IDS gradually decreases month by month before the onset of two-word utterances. In addition, in terms of parents' $F0$ variability, the analysis revealed that (1) the between-utterance variability is larger in IDS than in ADS both before and after the onset of two-word utterances, (2) the between-utterance variability in IDS is larger before the onset of two-word utterances than after it, (3) the within-utterance variability in IDS is the same as in ADS both before and after the onset of two-word utterances, and (4) the within-utterance variability in IDS is the same before and after the onset of two-word utterances.

These results show that, both for infants and parents, the $F0$ characteristics change significantly before and after the onset of two-word utterances. We suggest that mutual communication capability between infants and parents is closely related to this $F0$ change. In other words, one of the factors in the $F0$ change is the feasibility of communication between infant and parents.

It should be noted that the change does not occur suddenly. This change is gradual as can be seen in Figs. 1 and 2. This is probably because an infant's language development proceeds gradually, and the parents' changes in utterance proceed at a corresponding pace.

The input provided by a parent's utterance plays an important role in an infant's language development. In other words, parents' vocal interactions facilitate an infant's language development. The current results say nothing about this interaction, because the utterances of infants and parents were independently analyzed. Only the global tendency of the utterances is shown in this study. However, if we were to analyze a successive pair of utterances from infants and parents as McRoberts and Best (1997) did, some evidence of this interaction could be obtained. Although McRoberts and Best (1997) did not find any evidence of accommodation or imitation between successive utterances from infants and parents in terms of $F0$ mean, it might be found were we to use our huge and long-period utterance data. In addition, accommodation or imitation might be found in terms of SD. These points will be investigated in a future study.

Although this study provides clear results in relation to the onset of two-word utterances, we should be aware of certain limits for generalization. First, the results are based on a small number of participants. Moreover, the parents of Infants C and D are the same people. Because of this, the results might only reflect the particular individual characteristics of these participants rather than general characteristics. Second, the results are based on the free utterances of the participants in daily life. The utterances were not well controlled in terms of content, intention, environment, or other factors. This might result in fluctuations in the data as seen in Figs. 1 and 2. Such fluctuations might hide some true $F0$ characteristics. Finally, the data were divided into two parts and analyzed regarding the onset of two-word utterances. However, the onset of two-word utterances is arbitrarily set as a boundary. Although the results are fairly clear with this boundary and reasonable interpretations are possible, no statistical or mathematical evidence was provided for the use of the onset of two-word utterances as the boundary. Some other developmental point, such as the onset of one-word utterances, might provide a more valid boundary.

Further research is necessary to overcome these deficiencies. As for the first point mentioned above, increasing the number of utterances will enable us to perform a more precise and reliable analysis. For that purpose, we shall continue to develop the database to obtain more utterances. However, the maximum number of participants in the database is five, and this might be insufficient to ensure validity. Cross-sectional research should be conducted to confirm the results obtained in this work. As for the second point, utterances are categorized according to their content, grammatical type, and pragmatic type. If we analyzed each category, the fluctuation in $F0$ would be reduced and more reliable results would be obtained. As for the final point, a more precise statistical analysis is necessary to check the validity of the onset of two-word utterances as a boundary. The boundary can be identified by fitting the $F0$ change to a second-order function or a compound of flat-and decrease-first-order func-

tions. Another way would be to conduct short longitudinal observations with many participants over a given period with the boundary candidates. For example, half-monthly observations from 10 to 25 months would be sufficient to check the validity of using the onset of two-word utterances as the boundary.

Amano, S., Kato, K., and Kondo, T. (**2002**). "Development of Japanese infant speech database and speaking rate analysis," Proceedings of International Conference on spoken language processing, Vol. **1**, pp. 317–320.

Bennett, S. (**1983**). "A 3-year longitudinal study of school-aged children's fundamental frequencies," J. Speech Hear. Res. **26**, 137–142.

Burnham, D., Kitamura, C., and Vollmer-Conna, U. (**2002**). "What's new, pussycat? On talking to babies and animals," Science **296**, 1435.

Cooper, R. P., Abraham, J., Berman, S., and Staska, M. (**1997**). "The development of infants' preference for motherese," Infant Behav. Dev. **20**, 477–488.

Eguchi, S., and Hirsh, I. J. (**1969**). "Development of speech sounds in children," Acta Oto-Laryngol., Suppl. **257**, 1–51.

Fairbanks, G. (**1942**). "An acoustical study of the pitch of infant hunger wails," Child Dev. **13**, 227–232.

Fernald, A., and Kuhl, P. (**1987**). "Acoustic determinants of infant preference for motherese speech," Infant Behav. Dev. **10**, 279–293.

Fernald, A., and Simon, T. (**1984**). "Expanded intonation contours in mothers' speech to newborns," Dev. Psychol. **20**, 104–113.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., and Fukui, I. (**1989**). "A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants," J. Child Lang **16**, 477–501.

Grieser, D. L., and Kuhl, P. K. (**1988**). "Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese," Dev. Psychol. **24**, 14–20.

Hirsh-Pasek, K., Nelson, D. G. K., Jusczyk, P. W., Cassidy, K. W., Druss, B., and Kennedy, L. (**1987**). "Clauses are perceptual units for young infants," Cognition **26**, 269–286.

Hollien, H., Green, R., and Massey, K. (**1994**). "Longitudinal research on adolescent voice change in males," J. Acoust. Soc. Am. **96**, 2646–2654.

Hollien, H., Hollien, P. A., and de Jong, G. (**1997**). "Effects of three parameters on speaking fundamental frequency," J. Acoust. Soc. Am. **102**, 2984–2992.

Jacobson, J. L., Boersma, D. C., Fields, R. B., and Olson, K. L. (**1983**). "Paralinguistic features of adult speech to infants and small children," Child Dev. **54**, 436–442.

Kajikawa, S., Amano, S., and Kondo, T. (**2004**). "Speech overlap in Japanese mother-child conversations," J. Child Lang **31**, 215–230.

Katz, G. S., Cohn, J. F., and Moore, C. A. (**1996**). "A combination of vocal *F*0 dynamic and summary features discriminates between three pragmatic categories of infant-directed speech," Child Dev. **67**, 205–217.

Keating, P., and Buhr, R. (**1978**). "Fundamental frequency in the speech of infants and children," J. Acoust. Soc. Am. **63**, 567–571.

Kent, R. D. (**1976**). "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," J. Speech Hear. Res. **19**, 421–447.

Kent, R. D., and Bauer, H. R. (**1985**). "Vocalization of one-year-olds," J. Child Lang **12**, 491–526.

Kent, R. D., and Murray, A. D. (**1982**). "Acoustic features of infant vocalic utterances at 3, 6, and 9 months," J. Acoust. Soc. Am. **72**, 353–365.

Kitamura, C., Thanavishuth, C., Burnham, D., and Luksaneeyanawin, S. (**2002**). "Universality and specificity in infant-directed speech: Pitch modification as a function of infant age and sex in a tonal and non-tonal language," Infant Behav. Dev. **24**, 372–392.

Kubozono, H. (**1993**). *The Organization of Japanese Prosody* (Kuroshio, Tokyo).

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., and Lacerda, F. (**1997**). "Cross-language analysis of phonetic units in language addressed to infants," Science **277**, 684–686.

Laufer, M. Z., and Horii, Y. (**1977**). "Fundamental frequency characteristics of infant non-distress vocalization during the first twenty-four weeks," J. Child Lang **4**, 171–184.

Lee, S., Potamianos, A., and Narayanan, S. (**1999**). "Acoustic of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. **105**, 1455–1468.

Masataka, N. (**1992**). "Pitch characteristics of Japanese maternal speech to infants," J. Child Lang **19**, 213–223.

Masataka, N. (**2002**). "Pitch modification when interacting with elders: Japanese women with and without experience with infants," J. Child Lang **29**, 939–951.

McRoberts, G. W., and Best, C. T. (**1997**). "Accommodation in mean *F*0 during mother-infant and father-infant vocal interactions: a longitudinal case study," J. Child Lang **24**, 719–736.

Nakatani, T., and Irino, T. (**2002**). "Robust fundamental frequency estimation against background noise and spectral distortion," Proceedings of International Conference on spoken language processing, Vol. **3**, pp. 1733–1736.

Nakatani, T., and Irino, T. (**2004**). "Robust and accurate fundamental frequency estimation based on dominant harmonic components," J. Acoust. Soc. Am. **116**, 3690–3700.

Nakatani, T., Irino, T., and Zolfaghari, P. S. (**2003**). "Dominance spectrum based V/UV classification and *F*0 estimation," Proceedings of Eurospeech, pp. 2313–2316.

Niwano, K., and Sugai, K. (**2003**). "Pitch characteristics of speech during mother-infant and father-infant vocal interactions," Jpn. J. Spec. Educ. **40**, 663–674.

Papousek, M., Papousek, H., and Symmes, D. (**1991**). "The meanings of melodies in motherese in tone and stress languages," Infant Behav. Dev. **14**, 415–440.

Penman, R., Cross, T., Milgrom-Friedman, J., and Meares, R. (**1983**). "Mother's speech to prelingual infants: A pragmatic analysis," J. Child Lang **10**, 17–34.

Prescott, R. (**1975**). "Infant cry sound; developmental features," J. Acoust. Soc. Am. **57**, 1186–1191.

Ratner, N. B., and Pye, C. (**1984**). "Higher pitch in BT is not universal: acoustic evidence from Quiche Mayan," J. Child Lang **2**, 515–522.

Reissland, N. (**1998**). "The pitch of 'REAL' and 'Rhetorical' questions directed by a father to his daughter: A longitudinal case study," Infant Behav. Dev. **21**, 793–798.

Reissland, N. (**2000**). "The cradling bias in relation to pitch of maternal child-directed language," Br. J. Dev. Psychol. **18**, 179–186.

Reissland, N., Shepherd, J., and Cowie, L. (**2002**). "The melody of surprise: maternal surprise vocalizations during play with her infant," Infant Child Dev. **11**, 271–278.

Remick, H. (**1976**). "Maternal speech to children during language acquisition," in *Baby Talk and Infant Speech*, edited by W. von Raffler-Engel and Y. Lebrun (Swets and Zeitlinger, B. V., Amsterdam), pp. 223–233.

Robb, M. P., and Saxman, J. H. (**1985**). "Developmental trends in vocal fundamental frequency of young children," J. Speech Hear. Res. **28**, 421–427.

Robb, M. P., Saxman, J. H., and Grant, A. A. (**1989**). "Vocal fundamental frequency characteristics during the first two years of life," J. Acoust. Soc. Am. **85**, 1708–1717.

Sachs, J. (**1977**). "The adaptive significance of linguistic input to prelinguistic infants," in *Talking to Children*, edited by C. E. Snow and C. A. Ferguson (Cambridge University Press, Cambridge), pp. 51–61.

Shepard, W. C., and Lane, H. L. (**1968**). "Development of the prosodic features of infant vocalizing," J. Speech Hear. Res. **11**, 94–108.

Shute, B., and Wheldall, K. (**1989**). "Pitch alternations in British motherese: Some preliminary acoustic data," J. Child Lang **16**, 503–512.

Singh, L., Morgan, J. L., and Best, C. T. (**2002**). "Infants' listening preferences: Baby talk or happy talk?" Infancy **3**, 365–394.

Spence, M. J., and Freeman, M. S. (**1996**). "Newborn infants prefer the maternal low-pass filtered voice, but not the maternal whispered voice," Infant Behav. Dev. **19**, 199–212.

Stern, D. N., Spieker, S., Barnett, R. K., and MacKain, K. (**1983**). "The prosody of maternal speech: Infant age and context related changes," J. Child Lang **10**, 1–15.

Terasawa, R., Kakita, Y., and Hirano, M. (**1984**). "Simultaneous measurements of mean air flow rate, fundamental frequency and voice intensity," Onsei Gengo Igaku **25**, 189–207 (in Japanese).

Warren-Leubecker, A., and Bohannon III, J. N. (**1984**). "Intonation patterns in child-directed speech: Mother-father differences," Child Dev. **55**, 1379–1385.

Werker, J. F., and McLeod, P. J. (**1989**). "Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness," Can. J. Psychol. **43**, 230–246.

Loud speech over noise: Some spectral attributes, with gender differences

Sten Ternström
Department of Speech, Music and Hearing, Kungliga Tekniska Högskolan, Stockholm, Sweden

Mikael Bohman
Department of Speech, Music and Hearing, Kungliga Tekniska Högskolan, Stockholm, Sweden

Maria Södersten
Department of Logopedics and Phoniatrics, Karolinska University Hospital, Huddinge, Stockholm, Sweden

(Received 22 December 2004; revised 2 December 2005; accepted 4 December 2005)

In seeking an acoustic description of overloaded voice, simulated environmental noise was used to elicit loud speech. A total of 23 adults, 12 females and 11 males, read six passages of 90 s duration, over realistic noise presented over loudspeakers. The noise was canceled out, exposing the speech signal to analysis. Spectrum balance (SB) was defined as the level of the 2–6 kHz band relative to the 0.1–1 kHz band. SB averaged across many similar vowel segments became less negative with increasing sound pressure level (SPL), as described in the literature, but only at moderate SPL. At high SPL, SB exhibited a personal "saturation" point, above which the high-band level no longer increased faster than the overall SPL, or even stopped increasing altogether, on average at 90.3 dB (@30 cm) for females and 95.5 dB for males. Saturation occurred 6–8 dB below the personal maximum SPL, regardless of gender. The loudest productions were often characterized by a relative increase in low-frequency energy, apparently in a sharpened first formant. This suggests a change of vocal strategy when the high spectrum can rise no further. The progression of SB with SPL was characteristically different for individual subjects. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2161435]

PACS number(s): 43.70.Gr, 43.50.Qp, 43.72.Dv [AL] Pages: 1648–1665

I. INTRODUCTION

A. Earlier work on loud voice

To improve our understanding of the health risks that may be incurred by overly loud phonation, and why women are more at risk,[1] it is necessary to know in what ways the sound of the voice will change when the voice is driven near to its limits. Some acoustic features of shouting voice have been described in the literature, although we have not found any great number of studies. All such studies report that the fundamental frequency (F_0) and the sound pressure level (SPL) increase with effort, as would be expected. Most of the studies were done using only single vowels, words, or carrier phrases. Vilkman, Vintturi, and Alku[2] have recently published a series of studies, in which speakers of both genders produced /pa:p:a/ sequences over a very large range of effort levels. They found that some hitherto observed linearities, in relationships between input and output parameters in voice production, seem not to hold when extended to soft (<70 dB@40 cm) and loud (>90 dB) extremes of phonation. In technical systems, nonlinearities often arise when a nominal working range is exceeded, so, by analogy, this might be relevant to voice overload.

As for studies that explicitly concern voices over ambient noise, as in the work-environment, these have usually used noise presented over headphones, which can be problematic in terms of realism, and all of them were performed with only three to five levels of the SPL factor. Traunmüller and Eriksson[3] studied people speaking/shouting a model sentence outdoors over small and large distances. They measured many acoustic variables as a function of "vocal effort level" (VEL). This they defined *perceptually* as (the 2-logarithm of) the distance to the speaker as rated by listeners. This definition is elegant and convenient for several reasons. They found, for example, that the first formant frequency F_1 increases somewhat with F_0, to a degree that is compatible with larger jaw openings in loud voice, as found by Schulman.[4] Unfortunately, Traunmüller and Eriksson did not report calibrated SPL values for the voices. Neither did Frøkjær-Jensen,[5] who also found higher F_1 in shouting. An F_1/SPL relationship was quantified to +3.5 Hz/dB by Liénard and di Benedetto,[6] also using a distance paradigm. Rostolland[7,8] studied shouted voice (CVCV words), finding that speech quality deteriorates, vowels become longer in duration, and spectral energy tends to concentrate in a frequency range where hearing is more sensitive.

Huber *et al.*[9] measured F_0 and formant frequencies of [ɑ] vowels at three levels of vocal effort with no ambient noise. The sound levels were rather close, −5, 0, and +10 dB from comfortable, which, in view of our present data, would not bring the subjects into really effortful phonation. The effects found were not large. No SPL difference between adult males and females was reported.

Titze and Sundberg[10] described exhaustively the mechanisms that contribute to vocal sound pressure level, pointing out the significance of the excess lung pressure over the phonation threshold pressure, and found good agreement between predictions and experimental data. Although the voice

 0001-4966/2006/119(3)/1648/18/$22.50

SPL of some subjects reached 105 dB at 0.3 m, these subjects were professional tenors, and the finding is not directly applicable to speech or shouting.

It has been shown[3,11] that listeners are capable of accurately estimating vocal effort from recordings, regardless of the actual SPL at which the sound is presented. We may therefore ask if there exists some acoustic criterion for effortful and possibly risky phonation that would be independent of the perception of a listener or of the speaker. The purpose of the present study was to obtain basic data on loud and very loud voice, in normal speakers who are speaking or shouting so as to communicate in a fairly realistic noisy environment. In particular, we sought to study the evolution of the voice spectrum envelope toward very high effort levels. Given that women appear to be more prone to voice problems than men,[1] we sought also to compare female and male voice at high effort levels.

B. Hypothesis

If the spectral energy distribution of a sound on the whole decreases toward higher frequencies, we may say that the spectrum slope is negative, e.g., −12 dB/octave. A negative spectrum slope is typical of practically all natural sounds, including the human voice. It is well known that with increasing vocal effort, the voice spectrum slope toward higher frequencies on the whole becomes less negative.[12] Henceforth in this paper we will say that the spectrum slope *increases* when it goes from negative to less negative and thus becomes *less* steep. This is not conventional, but in our view it is logical, when discussing the voice.

For the widely studied LF model of glottal flow,[13] it has been shown[14] (a) that the SPL depends on the amplitude of the negative peaks in the first derivative of flow at glottal closure (the maximum flow declination rate), and (b) that the source spectrum slope increases as the duration of the return phase to zero flow derivative approaches zero. Although the real voice source is more complex than the LF model, a short return phase can be associated with abruptness of closure, and it also affects the second derivative of glottal flow. If the folds collide with high velocity and meet at the same instant along their full length, the abruptness is maximal. If the folds first meet partially and then close gradually along their length, then there is less abrupt closure; likewise if their neutral position is somewhat abducted.

We hypothesize that, above some individual effort level, the source spectrum slope will stop increasing with effort. In other words, as the voice becomes louder, the spectrum slope also rises, but probably only up to a point, that is, up to a certain voice SPL. The voice can be pushed above this SPL and become even louder, but its spectral balance cannot continue shifting toward high frequencies. We shall call this hypothetical effect "spectrum slope saturation," as a working name only.

The theoretical rationale for source spectrum slope saturation is that high-frequency energy in the voice source is generated by glottal flow changes, which cannot be infinitely rapid. In the limit, if the time taken to shut off the glottal flow goes to zero, the flow spectrum would approximate that of an ideal sawtooth waveform, which has a slope of −6 dB/octave. The flow derivative, corresponding to a pressure signal, would then be a train of Dirac pulses, which has a flat spectrum, 0 dB/octave. Toward high frequencies, then, the source spectrum slope must always be negative, in any plausible physical scenario.

Practical limitations in the biomechanical domain should further restrict the spectrum slope and/or the bandwidth of the glottal source. For example, there is surely a maximum vibratory amplitude of the vocal folds: they cannot stretch indefinitely, and they are restrained medially by the surrounding tissues. A maximum vibratory amplitude would indirectly constrain the source spectrum slope, in that it limits the magnitudes also of the flow derivatives. If maximum vibratory amplitude is reached at a flow that is less than maximum, then increasing the subglottal pressure will still make the sound louder, due to a larger flow pulse amplitude, but this should only scale the waveform amplitude without affecting the spectrum slope.

II. ACQUISITION

A. Choice of method

How may one determine by experiment if the spectrum slope does in fact "saturate"? It would seem logical to ask subjects to perform sustained vowels with increasing effort, and study the evolution of the voice spectrum. For an untrained speaker, however, this could be an unrealistic and difficult, singing-like task. Sounds thus produced in the laboratory could well be rather different from those produced by speaking or shouting in authentic work environments. A task using running speech seems desirable.

In an earlier paper,[15] a method was described for recording speaker subjects in simulated environmental noise presented over loudspeakers. The noise can provoke the subjects to raise their vocal effort to near maximum, but it also contaminates the recording. Through signal processing, the noise can be attenuated by up to 60 dB, and the airborne voice of the subjects can be recovered with excellent accuracy. In the present investigation, this method was used to invoke realistic loud speech, for the purpose of trying to characterize such speech acoustically. No attempt was made to differentiate the influence of the different noise types on voice usage.

Assuming that the spectrum slope variation with vocal effort is attributable to the glottal source, then the principal method for analysis would be inverse filtering. At high F_0, unfortunately, inverse filtering is difficult to do manually, and quite precarious to do automatically. Inverse filtering requires some *a priori* assumption about the overall source spectrum, which in this case is the object of inquiry.

The long time average spectrum (LTAS) may be used to estimate the degree of vocal press, but with running speech it takes at least 30 s for the average to settle down.[16] Even then it will be dominated by the loudest phonations in that interval. Here, we are not interested in a time average, but rather in obtaining averages of many similar vowel segments produced at specific sound levels.

If instead we take running speech and examine each voiced phoneme in turn, we will find samples of vowels

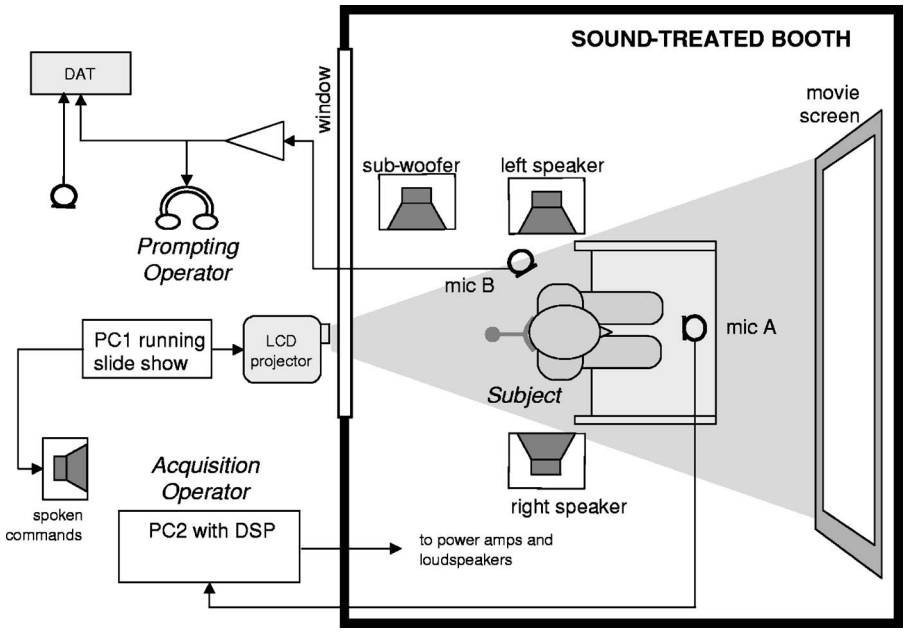


FIG. 1. The layout of the acquisition setup.

produced at many different effort levels, thanks to prosody and other variations in running speech. Under the assumption that voice SPL is sufficiently correlated with vocal effort, individual segments with similar vowels can be sorted by increasing SPL, and the evolution of their spectrum slope toward high effort can be studied. This is the approach we chose. However, it requires a considerable degree of automation in order to be practical for recordings of running speech.

B. Acquisition setup

1. Topology and equipment

The layout of the acquisition setup is shown in Fig. 1. The recording booth had the internal dimensions 3.30 $\times 3.05 \times 2.15$ m^3. It was sound-treated but not anechoic. The subject was seated in a chair with armrests and an adjustable neck support. Microphone A was a highly directional hypercardioid back electret microphone (AOI Electronics ECM-1014), placed 30 cm in front of and slightly below the subject's mouth. The subject's body below the neck was screened from microphone A, using a small wooden kiosk fitted with absorbent panels. This reduced the influence of body movements on the acoustic channel from the loudspeakers to microphone A (Fig. 2). There were two loudspeakers (Audio Pro A4-14 Mk II) on stands to either side of the subject, at a distance of approximately 0.6 m. The noise from the loudspeakers reached the subject's ears directly, but would arrive at microphone A only after some diffraction and/or damped reflections. This screening, combined with the directivity of the microphone, served to reduce the level of the noise at microphone A by some 15 dB, compared to the noise level at the subject's ears. On the floor was an extra sub-woofer loudspeaker, custom built by author MB. It had a dedicated equalizer that was carefully adjusted, together with the exact position of the sub-woofer, to give as smooth a low-frequency response at the subject's head as was possible in the small booth.

The sound near the subject's left ear was monitored using microphone B (omnidirectional Brüel and Kjær model 4003). An omnidirectional microphone must be used here, in order to obtain the same airborne voice-to-noise ratio as that experienced by the subject.

The experimental procedure was sequenced by a slideshow presentation running on computer 1, under the control of an operator outside the booth. The slideshow was projected through the booth window onto a large screen (2.0 $\times 1.5$ m^2) inside the booth. The slides instructed the subjects, and displayed the texts for reading. The slides also contained audio clips with spoken instructions for the procedural sequence to the operators outside the booth, such as "load noise 3," "start tape," etc.

In addition, the slides carried a vertical bar indicator on which the slideshow operator would point with the cursor, to indicate whether or not the subject's voice could be heard

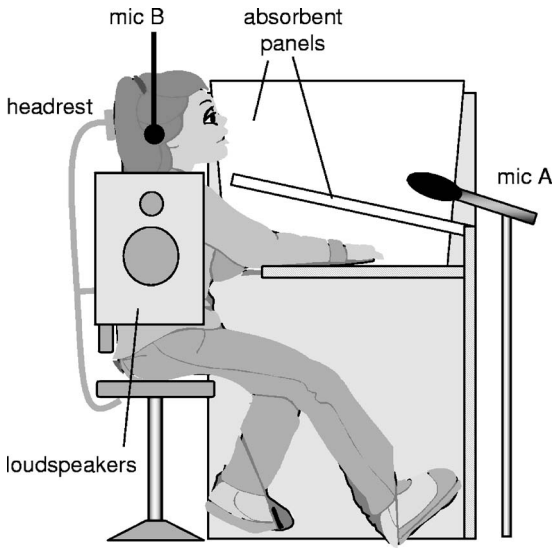


FIG. 2. The acoustic screening "kiosk" and placement of the side loudspeakers, the subject, and the microphones.

TABLE I. The site noises used in the six experimental conditions.

Condition	Noise type	Level (dBA)	Suggested situation and audience
Quiet 1	None	<30	Quiet classroom. Children about to go by bus to a theme park.
Soft noise	Stationary white noise, low-passed at 200 Hz, −6 dB/oct	Ramped down 78–70	Outside a ventilation engine room. Adults about to take a technical tour.
Preschool	Preschool babble recorded at teacher's ears	74	Noisy classroom. Children about to go by bus to a theme park.
Disco	Pub with loud music	87	A rowdy night out. Young adults about to have dinner and a show.
Loud noise	Stationary white noise, low-passed at 200 Hz, −6 dB/oct	Ramped up 78–85	Inside a ventilation engine room. Adults about to take a technical tour.
Quiet 2	None	<30	Quiet classroom. Children about to go by bus to a theme park.

through the noise. The bar was colored with a gradient, from green at the top ("can hear you") to red at the bottom ("can't hear you"). The purpose of this visual feedback to the subject was only to let the operator coax the subject a little in case he or she were perceived to be making a less than sincere effort when reading. The presence of this "audibility meter" implies that the recordings do not reliably represent the levels at which the subjects themselves would choose to speak in the given types of noise. However, the purpose of this experiment was only to investigate some acoustical features of loud running speech.

2. Signal paths

The signal from microphone A was amplified by a Symetrix SX202 preamplifier and fed to one A/D input of the signal processing board fitted in computer 2. The board was a Loughborough Sound Images PC/C32 DSP board, clocked at 60 MHz and equipped with true 16-bit delta-sigma A/D converters, and integral anti-aliasing filters. The sampling rate was 22050 Hz throughout. The DSP board ran a signal model built using interactive DSP software (Aladdin Interactive DSP 1.30 by Hitech Development AB). The model played back a sound file with the particular site noise, and simultaneously recorded the signal from microphone A. It also recorded the position of the cursor of the "audibility meter," using a modified computer mouse, but that information was not used.

The signal from microphone B was heard by the slide-show operator outside the booth. A DAT recorder ran continuously, recording on one channel the signal from microphone B, and on the other channel the sounds *outside* the recording booth, including the sequence of events and the conversation of the operators. The DAT recordings served only to document the progress of the experiment, and were not part of the material to be analyzed.

C. Noise conditions

The noise conditions are listed in Table I. The order of presentation was fixed. There were two Quiet conditions with no ambient noise, presented first and last. There were two noise types recorded on location: preschool babble recorded at a teacher's ears, with an equivalent level of 74 dBA; and loud music with heavy bass recorded at a table in disco pub (87 dBA). There were also two conditions of white noise, low-pass filtered at 200 Hz and −6 dB/octave. In the latter two conditions, the sound level was very slowly ramped during the task over an 8 dB range. The level change was so slow as to be almost imperceptible. This was done in order to obtain samples of reading at a more continuous range of voice sound levels. In a previous study,[15] we had found that subjects tended to stay rather close to an effort level chosen for each condition, thereby clustering around a few discrete effort levels.

It must be emphasized that the different noise types were only a device for eliciting a wide range of vocal efforts. They were not experimental conditions in the usual sense and their effect was not tested. In other words, the effect of the noise type on the speech is *not* a topic of this investigation, even though the nature of the experiment seems to suggest otherwise. In the previous study,[15] no effect had been found of the noise *type* on nonspectral voice measures, nor on the spectral centroid; only noise SPL made any difference. The different noise types were retained because they simulated natural noisy environments.

D. Subjects

The subjects were 12 females (mean age 44 years, range 30–54), and 11 males (mean age 34 years, range 20–53). Of the females, ten were preschool teachers and two were academics; nine were native speakers of Swedish. Of the males, one was preschool teacher, and the rest were academics; ten were native speakers of Swedish. The subjects were considered healthy. None had sought medical help for voice problems or had a diagnosed hearing loss. However, four of the female preschool teacher subjects had subjective symptoms of a slight hearing problem that one subject related to many years of work in a noisy day-care working environment. One male subject had symptoms of a hearing problem that he related to his background as a pop musician, and one male suffered from tinnitus. Two of the female preschool teacher

FIG. 3. The sequence of events for each noise condition.

subjects reported signs of vocal fatigue during work. None of the subjects had received extensive voice training, and none had an upper respiratory infection the day of the recording. Two subjects, one female and one male, each smoked about 10 cigarettes per day.

Approval from the Ethical Committee of Huddinge University Hospital was obtained prior to data collection. The subjects who had to take time out from work and travel to KTH to participate were compensated with the amount of SEK 200. Subjects recruited from within KTH were not compensated.

E. Procedure

The subjects had been provided in advance with a handout describing the experiment and containing the texts that they would be asked to read. On arrival, information was recorded about their sex, age, occupation, vocal experience, known hearing loss, smoker status, and general health condition.

The subject was seated in the chair and its head rest was adjusted. The nominal mouth-to-microphone distance was 0.30 m. For each subject, the actual microphone distance was measured and compensated for in the subsequent level calibrations. The estimation noise (see Sec. II G) was played over the speakers, and its level in dBA was measured at the position of microphone B.

The booth was then closed and the slide show was started. The first slides explained the general procedure, including the "audibility meter," and emphasized the importance of trying to remain still. There was a short trial run of reading for about one minute to let the subject try the task first. Then, for each of the six noise conditions, the following procedure was followed (Fig. 3).

1. The estimation noise was played for 8 s.
2. The site noise was played, and the subject started to read out loud. The signals from microphone A and from the audibility meter were recorded.
3. The site noise and recording were stopped as soon as the subject had finished reading, after 80–110 s.
4. The questionnaire was given.

The order of the conditions was not randomized but fixed as shown in Table I. It seemed reasonable to save the loudest task for second to last, and then to finish with a replication of the Quiet condition. The entire procedure took 22–30 min per subject. The recordings took place over three days, and resulted in 138 takes (23 subjects $\times$ 6 site noises), each of 85–110 s duration. All takes were useable.

F. Text material

The reading material for each condition was a text of instructional character, designed for a fictional audience in a situation matching that of the site noise. There were three texts (see the "situations" column in Table I), in which the content and structure were very similar, but the details were different. These variations were intended to keep the subject alert and motivated. While reading, the subject would discover: *I am set in some noisy environment, cheerfully instructing a group of people who are about to spend some time together. There are several contingencies they must be prepared for, and several rules of conduct that they must follow, and I am responsible for getting that information across*. Each complete text fit on one slide. A washed-out background photograph of a day-care center or an industrial machinery site was included as an unobtrusive "wallpaper" on the text slides where appropriate. The texts took 80–110 s to read.

G. Noise cancellation

The next stage of the data acquisition was the noise removal, which is an off-line data processing operation. This time, the noise cancellation processing was more ambitious than in our previous study. Its details are described elsewhere.[17] Briefly, the idea is to create a digital filter that emulates precisely the transfer characteristic of the entire acoustical path of the site noise, from the computer's D/A converter, via the loudspeakers, the recording booth, and the acoustic screens, through the microphone and back through the computer's A/D. Before each condition, with the subject seated in place but silent, a wide-band estimation noise is played through the system and re-recorded. By comparing this re-recorded noise to that which was sent to the loudspeakers, a so-called channel estimation can be performed. Channel estimation is a textbook signal processing procedure that produces a digital filter that emulates a given time-invariant acoustic channel to any desired accuracy. Here, we used a filter of 4096 frequency points, corresponding to a resolution of 5.4 Hz from 0 to 11025 Hz. The impulse response of the adaptive filter was 0.186 s long. The reverberation time T60 of the recording booth was on the order of 0.2 s. In addition to the adaptive filter, a fixed bandpass filter 80 Hz–10 kHz was applied, so that the lowest room resonances (and much of the bass in the disco condition) did not need to be canceled.

After the channel estimation, the filter adaptation is inhibited, so that the filter characteristic is held constant. Then, by playing only the site noise through the estimated filter, a signal is obtained which closely resembles that coming

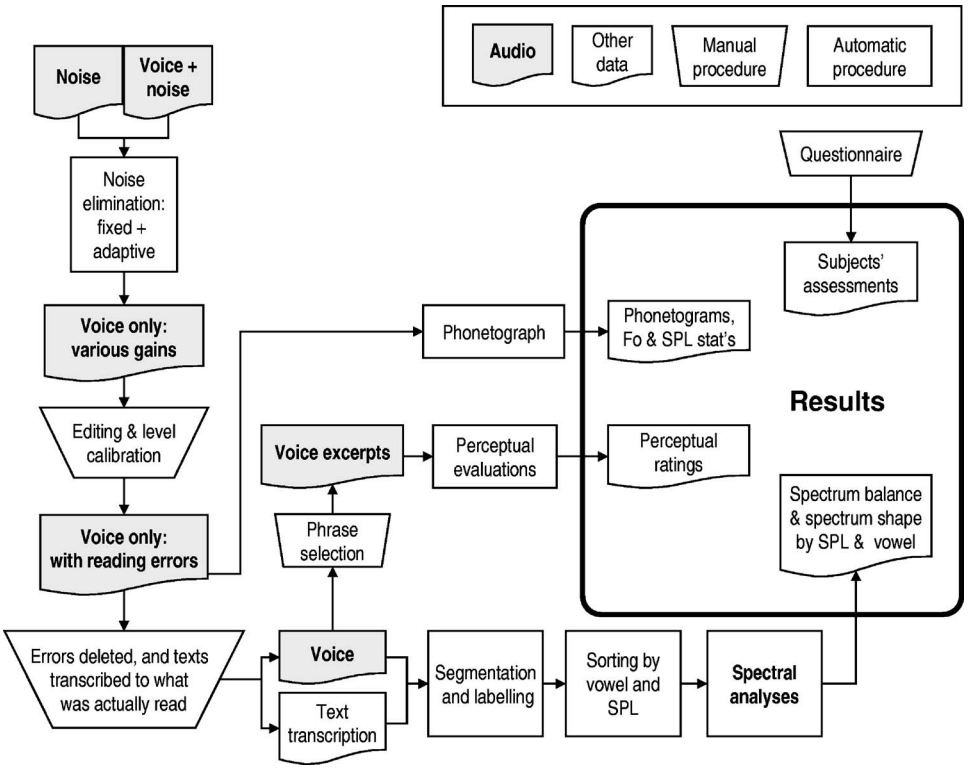


FIG. 4. Flow chart for the four methods of analysis. The spectral analysis is reported in this paper, while the other analyses are reported in Södersten *et al.* (Ref. 18).

through the acquisition microphone—provided that the subject is silent and sits still. When the subject starts speaking, the voice signal alone can be recovered by subtracting the signal coming through the estimated filter from the signal coming through the microphone.

Since each new subject will affect the acoustic channel differently, and also will tend to move and shift in the chair, the filter must be reestimated for each subject and for each site noise condition. As a further refinement, the system detects and exploits brief pauses in the running speech to update the filter adaptation, thereby to accommodate minor changes in the acoustic channel during the reading. This improvement virtually eliminated the problems we had in a pilot study, when the acoustic channel would be perturbed by small body movements of the subject. The end result is a reduction of the site noise level of typically 60 dB, and in the best cases up to 65 dB, relative to the noise level heard by the subject and by microphone B. A modern personal computer is likely to run the complete signal processing faster than real time, using a MATLAB implementation.

H. Editing and level calibration

Once the site noise had been removed from the recordings, they were processed by a script that calibrated the signal level, compensating for the actual microphone distance, and removed initial and final silences. A working copy of the entire set of "decontaminated" and calibrated recordings was made at this point, so that a phonetographic analysis could proceed in parallel. The recordings were then edited to remove coughs and redundant repetitions (there were not many). A transcription of the nominal Swedish text (not phonetic symbols) was made for each take, corrected for reading mistakes so that it matched the syllables that each subject actually had said. For each take, we now had a noise-free speech audio file with a matching text transcription.

III. ANALYSIS

The voice production of the subjects was analyzed in four different ways, namely (1) subjective self-rating by questionnaire, (2) computer-based phonetography, (3) perceptual ratings by expert listeners, and (4) spectrum analysis (Fig. 4). The outcomes of the first three modes of analysis are reported in a companion paper.[18] Only the spectrum analysis results are reported here.

A. Vowel group categorization

We first tried simply to pool *all* voiced segments, but soon found that the SPL differences between different vowels produced at similar effort were so large that using the SPL as the only sorting criterion gave far too much variance in the data. A loud [u] vowel, for example, may have a steeper spectrum, and hence a more negative spectrum slope value, than a soft [ɑ] vowel. We also tried short time averaged spectra of various lengths (0.5–10 s) in the hope that some prosodic SPL variation would still be visible within conditions, but this was not successful. In order to chart the spectrum evolution with SPL, it proved necessary to analyze the individual vowels separately.

Fortunately, we have access to an automatic segmentation program that is based on speech recognition techniques.[19] Using the text transcriptions of the running speech as a collateral, this program produced a machine-readable list of labels, identifying the sample-by-sample locations and durations of each individual phoneme in the

speech file. The phoneme identification of the automatic labeling was verified by manual inspection and adjusted where necessary.

After some experimentation, we settled for selecting only vowels from three distinct groups, containing the long and short versions of the Swedish cardinal vowels [i][y][e], [u][o], and [a]. With the six conditions pooled, this gave a total of a few hundred similar vowel segments per subject, spread across the voice effort range from low or moderate (Quiet conditions) to maximum (Loud Noise condition). The texts read by each subject contained a total of 309 [i]-like vowels, 514 [u]-like vowels, and 548 [a]-like vowels.

The labeling algorithm locates the expected phonemes using a best sequential match to the order that is given in the text transcription. The orthographic transcription is converted under a rule system to phonemes, which are then subjected to a Viterbi forced alignment to the audio waveform. This means that a particular segment will be labeled as an [a] vowel, for example, if its formant structure fits reasonably well where an [a] is expected in the context of the transcription. It also means that in matching to vowel categories, considerable variation is tolerated. Several rather different renditions of a particular vowel would still be categorized as that vowel which was given in the text. This property of the labeling algorithm is actually useful, since we may expect vowels to be severely modified and phonetically reduced toward high vocal effort.[3,5,8] For the purposes of this analysis, such vowel reduction would operate in our favor, making it less critical to find vowels that are very similar.

B. Spectrum envelopes

In order to obtain some overview of the evolution of the spectrum envelope from low effort to high effort, a MATLAB script was written that created a composite view of all spectrum sections of a specified cardinal vowel group from a specified subject. The script performed the following operations. The desired set of vowel segments (containing, for example, all [u]-like vowels produced by subject 12 in all conditions) was compiled from the speech files. Within each vowel segment, the maximum point of the amplitude contour was located. A modified Hanning window of up to 50 ms duration ($<$50 ms if the vowel was shorter) was time-aligned with its midpoint as near as possible to the time of maximum amplitude. (The Hanning window was computed as the tenth root of the standard Hanning window. This gives a more square window, which attenuates the signal by less than 1 dB, rather than 4.3 dB for the normal Hanning window.) The windowed waveform was then passed through a third-octave filter bank, giving a logarithmic frequency scale with intentionally modest resolution. Each third-octave band was implemented using third-order Butterworth bandpass filters with skirts of 18 dB/octave. The resulting filter-bank spectra, one per 50-ms segment, were sorted by ascending total SPL into 1-dB bins. Within each third-octave band, a running median filter spanning five bins was applied (more details below), producing one filter-bank spectrum section for every 1-dB step in SPL. In Fig. 5, every third such spectrum section is plotted, in a stacked fashion, so as to produce one spectrum envelope for each 3 dB interval in SPL. In total, 69 such stacks were made (3 vowel groups $\times$ 23 subjects).

Some observations made from these plots will be mentioned in the results section. Although interesting to look at, we found it difficult to formulate any unambiguous measure of spectrum slope from this representation. At the highest SPL's, about half of the spectrum stack plots exhibited a clear peaking effect in the F1 region. This is an interesting observation in itself, but the shifting F1 peak made it difficult to establish a fixed pivot point for the spectrum slope.

Spectrum slope measures are elusive, and several candidates are found in the literature. Titze and Sundberg[10] performed inverse filtering and chose the difference in dB between partials 2 and 1 (one octave apart) as an estimate of the slope at low frequencies. Also after inverse filtering, Traunmüller and Eriksson[3] chose to use the energy ratio between the fundamental and the sum of all other partials. For phonetogram displays, Pabon[20] compared three spectrum slope measures, including the above two but without inverse filtering, and the time-domain crest factor, without arriving at a definitive choice.

C. Spectrum balance

Statistical techniques exist for factorizing compound effects, e.g., on spectrum envelope shapes such as these. However, since the objective here was simply to investigate the existence of a saturation effect in the high-frequency part of the spectrum, we chose instead to restate our hypothesis in terms of a measure of spectrum balance, which is much simpler and does not require inverse filtering. Guided by the appearance of the spectrum envelope plots, we divided the voice spectrum into a low band and a high band using two bandpass filters: 100–1000 Hz and 2–6 kHz, both 18 dB/octave (Fig. 6, dashed curves). The *spectrum balance* SB was defined as the level difference between the high-frequency band and the low-frequency band, $L_{\text{high}} - L_{\text{low}}$. This value will usually be negative, since the low band usually dominates the voice spectrum. Conveniently, SB typically increases (becomes less negative) as SPL increases.

A MATLAB script was written to measure and plot $\text{SB}_{\text{segment}}$ against total SPL, for all vowel segments pooled from all noise conditions. An example is shown in Fig. 7(a), for the pooled {a|ɑ} vowels of one male subject. The scatter in $\text{SB}_{\text{segment}}$ from segment to segment was always substantial, even with the vowels grouped as described above. Smoothing was performed using a median filter, as follows. The median filtered SB, $\text{SB}_{\text{median}}$, was computed by passing the SB values of the SPL-sorted vowel segments sequentially through a sliding median filter, 5 dB wide on the SPL axis, and moved 1 dB at a time. In other words, this median filter would output one $\text{SB}_{\text{median}}$ value for each 1 dB step in SPL. The advantages of using a median filter rather than a running average are that outlier points (which would sometimes appear due to errors in the vowel segmentation algorithm) are entirely and appropriately excluded; while abrupt but consistent changes in SB as a function of SPL are not smoothed. Since the vowel segments were extracted from 6 $\times$ 90 s of running speech per subject, the median filter would span

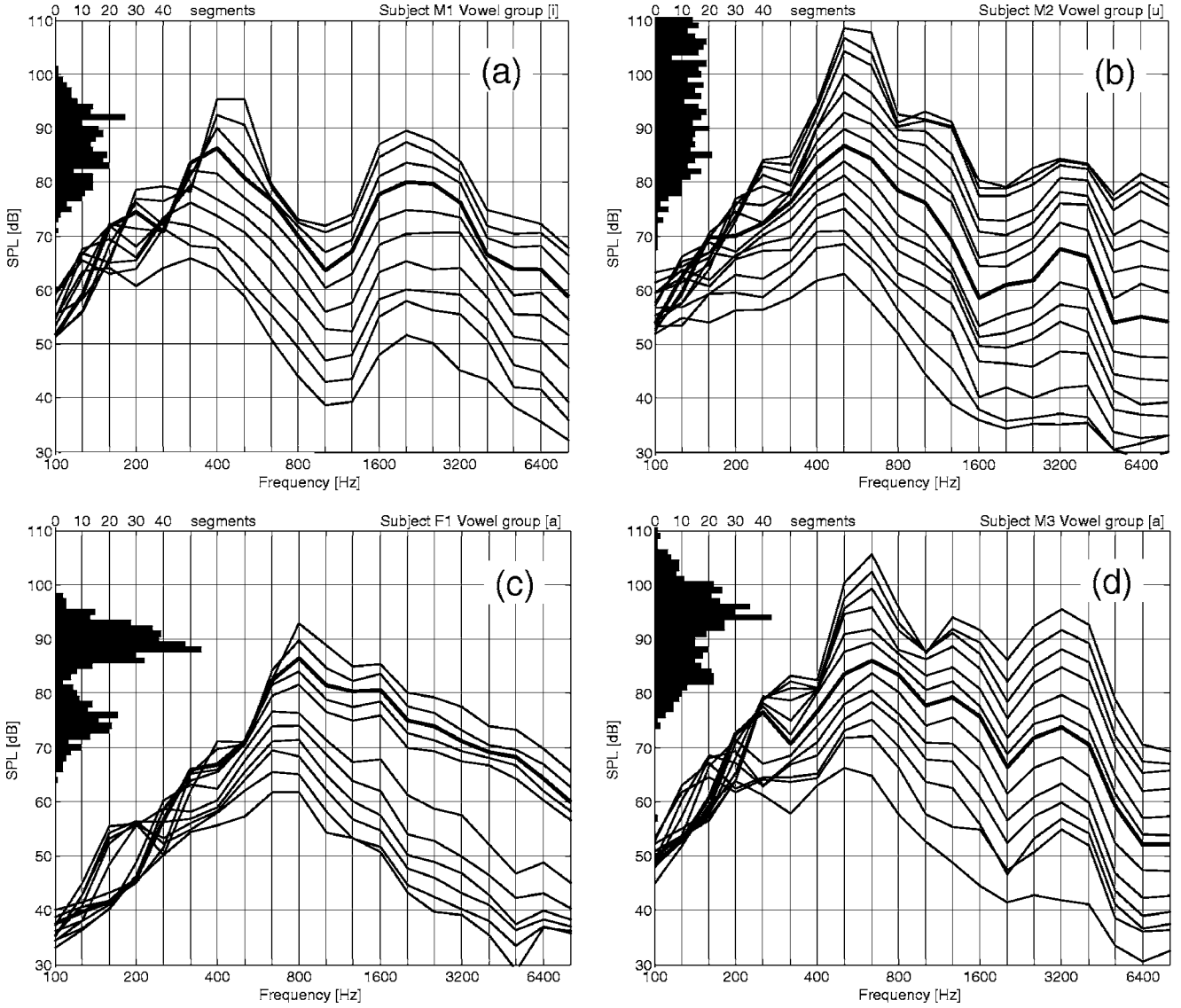


FIG. 5. Examples of spectrum shape changing with SPL. The horizontal axis is frequency in Hertz, and the vertical axis is third-octave band level in dB. Note that the logarithmic division of the frequency scale imparts a +3 dB/octave high-frequency bias to these figures, as compared to linear spectrum sections. Each part (a)–(d) shows a third-octave-band analysis of several hundred vowel segments from one subject and one vowel group, sorted by increasing SPL into 1-dB bins. The number of segments in each bin is shown in the histogram along the SPL axis. The SPL increment between adjacent third-octave spectra is 3 dB. For each spectrum, the level shown in each third-octave band is the median value for that band taken over five adjacent bins. The bolder line is the spectrum with median filters centered on the 90-dB bin. The four examples are discussed in Sec. IV A of the text.

across very many points at midrange SPL's and fewer points at the very low and very high extremes of SPL. Some examples of the distributions of vowel segments over SPL can be seen in Fig. 5, in the histograms that are plotted along the SPL axis.

Figure 7(b) shows only the SB_{median} curves for all three vowel groups of one subject. The differences in SB_{median} were consistent with the vowels' characteristic spectrum envelopes: the closed vowels {u|o} have relatively less high-frequency energy than do {a|ɑ}, while the vowels {i|y|e} have more. All subjects exhibited this same general difference in SB_{median} between the vowel groups.

Our revised hypothesis was that, for each speaker and for each vowel group, the SB would stop increasing above some high voice SPL. This "saturation" SPL, if found, will be called SPL_{sat}, and the corresponding spectrum balance value will be called SB_{sat}. To test the hypothesis on such very scattered data is not trivial, especially since the interesting part is near the top end, where the data points start to thin out, and the output of the median filter therefore becomes less stable. The data points could be smoothed by some other method, for example by fitting polynomials or passing the points through a running average filter. However, as with the median filter, the results might vary with the polynomial order or the filter length. A nonparametric smoothing method was sought, so that the test for saturation would be independent of any smoothing parameters.

Therefore, we chose to test for the presence of saturation as follows. Within each vowel group, the vowel segments were first sorted according to increasing total SPL. Then, for each vowel segment in turn, thus taking higher and higher SPL's into account, the energies in the low and high bands were *accumulated* separately. Thus, two data series were produced describing the total energy seen in all segments, for

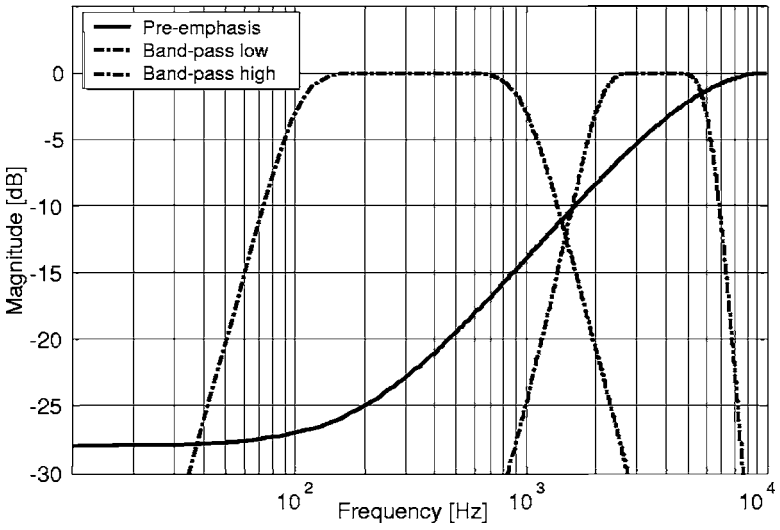


FIG. 6. The two dashed curves show the frequency response of the third-order low and high bandpass filters used for measuring the spectrum balance SB. The solid curve shows the frequency response of the high-frequency emphasis filter used for an alternative estimate of saturation SPL: +6 dB/octave from 200 Hz to 5 kHz. The choice of filter characteristic had little effect on the obtained saturation SPL.

either band, as a function of SPL. The high-to-low band ratio of the accumulated energies was named the Sorted Energy Ratio (SER), given in decibels. This method was inspired by Schroeder's reverse integration method for computing level decay envelopes of room acoustic impulse responses.[21] Because there is usually less energy in the high band, the SER in decibels was usually negative, although there were a few exceptions on {i|y|e} vowels of female speakers. Thanks to the large number of data points, the SER becomes very stable toward the end of each data series. Also, since the vowel segments were sorted by total SPL to start with, we can be very confident that, if the SER stops increasing with SPL when additional segments are accumulated, then the spectrum balance SB has indeed saturated.

(An analogy with balancing scales can illustrate how the SER works. To determine the mass ratio of a pair of items, we put them in the two trays of the scales and find the pivot point on the beam for equilibrium. When a new pair is added to the trays, the equilibrium shifts a little, unless the mass ratio of the new pair exactly equals the ratio of the masses already in the dishes. With hundreds of pairs already in the trays, each new pair will shift the equilibrium point by a very small amount. The SB of one vowel segment corresponds to the mass ratio of a pair of items. The SER corresponds to the ratio of the accumulated masses in the two trays, and to the position of the pivot point at equilibrium.)

To determine objectively whether or not SER (and hence SB) did stop increasing with SPL, an algorithm that implemented a test for saturation was coded in MATLAB. The algorithm can be summarized as follows (Fig. 8).

1. From the plot of SER vs SPL, find the vowel segment that incurred the maximum SER, and note its SPL.
2. If there are fewer than ten vowel segments with higher SPL than this point, conclude that the SB did not saturate.
3. Search the vowel segments sequentially from 80 dB SPL and upwards, for the first segment whose SER is within 0.5 dB of the maximum SER. This tolerance margin served to overcome small random fluctuations in the SER.
4. Select the SPL of the segment found in step 3 as SPL_{sat}.

Because the signal energy grows as a power function of the SPL, the SER is more sensitive to *increases* than to *decreases* in the level of either of the frequency bands. This implies that the above procedure is a conservative test for saturation. At high SPL, with many accumulated vowel segments, even a drastic drop in the *absolute* level of the high-frequency band will show up only gradually in the SER. So long as SPL and SB are increasing, however, the value of SER is quite similar to that of SB_{median}, for a given segment SPL.

For an independent assessment of saturation, SPL_{sat} was also determined by a completely different method, as follows. For each vowel segment, the high- and low-frequency bands were separated using a simple first-order preemphasis, rather than with third-order bandpass filters. The high band signal was obtained using a first-order high-frequency emphasis filter with a characteristic due to Fant, +6 dB/octave from 200 Hz to 5 kHz, and unity gain above 5 kHz (Fig. 6, solid curve). The level from this filter was compared to the level of the unfiltered signal. This method should result in SB values that are too low, because the total SPL is always larger than or equal to that in a subband, and because the high-band gain is less than unity below 5 kHz. The smoothing was performed by fitting tenth-order polynomials to the SB points. The resulting smoothed SB-vs-SPL plots were inspected visually for saturation. This validation procedure was carried out before the results of the SER-based procedure were known.

The two procedures were both iterated 69 times [3 vowel groups $\times$ (11♂ + 12♀) subjects]. When averaged separately over the male and female speakers, the two methods gave means of SPL_{sat} that agreed to within about 1 dB for all vowel groups. The SB_{sat} values from the second method were indeed lower, but since they were known to be incorrect, they were discarded.

IV. RESULTS

A. Spectrum envelopes

Figure 5 shows four examples, out of a total of 69, of the third-octave-band voice spectrum stacks, by subject and vowel group. Each stack represents several hundred vowel

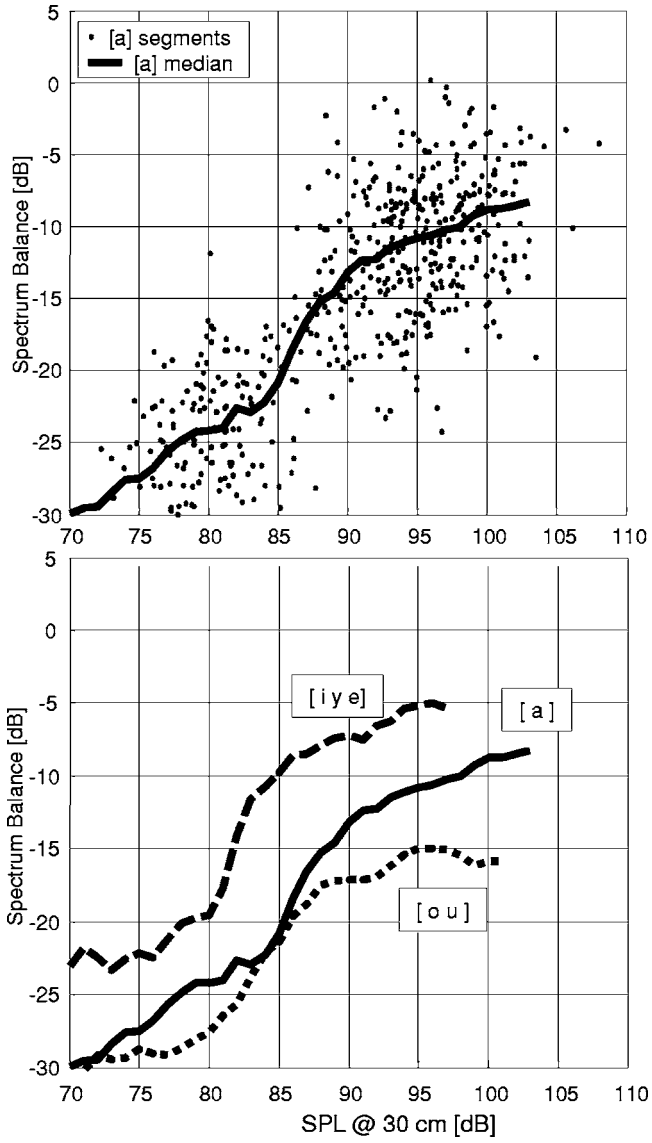


FIG. 7. Example of the unfiltered spectrum balance SB (vertical axis) versus SPL at 30 cm (horizontal axis) for a male subject. Data from all six noise conditions, from moderate to very high voice effort, are pooled. Upper graph (a): each dot denotes one data point from one 50-ms vowel segment from the group of [a] vowels. The solid line is the median. Lower graph (b): median SB curves for all three vowel groups of the same subject. The horizontal and vertical offsets for the [i] vowel group are due to the lower SPL and greater high-frequency energy than for [a]; and conversely, the lower SB values for the [u] vowel group are due to the more closed vowels having less high-frequency energy. Note how the median curves change with the overall SPL. The [i] and [u] curves level out at high SPL, while the [a] curve in this example does not. The median values in this figure, and elsewhere in this article, were computed from points under a sliding 5-dB SPL window that was moved in steps of 1 dB on the SPL axis.

segments, pooled from the six noise conditions. In (a), (b), and (c), the region above 2 kHz climbs rapidly in level from low to moderate SPL's, but then less rapidly, in (b) even reaching a ceiling at high SPL's. This would support the spectrum saturation hypothesis. Furthermore, the F1 peak often became quite narrow and high (a, b, d). As mentioned in Sec. III B, this can result in a reversal of the familiar rising trend in spectrum slope that is seen in the moderate SPL range. An upward shift in the frequency of F1, reported by others for loud voice, is also noticeable in (a), (b), and (c). The formant patterns manifest in these examples were representative of the vowel groups: (a) two peaks for {i|y|e}; (b) an F2 "shoulder" at about 1 kHz in {u|o}; and (c, d) an F2 "hill" at 1.2–1.6 kHz for {a|ɑ}. There was no obvious difference between genders in these plots, other than somewhat lower formant frequencies in males, as expected, and a higher average SPL in males, as reported earlier.[18]

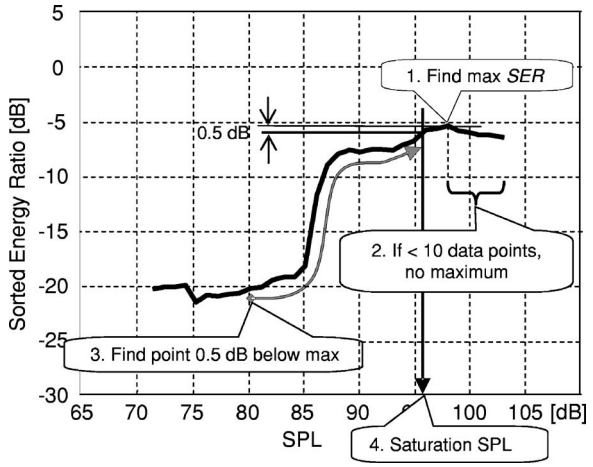


FIG. 8. The algorithm used for testing the presence of saturation via the sorted energy ratio SER (see text). For clarity, the solid line is a sketch of a possible graph of SER. (1) Find the maximum SER, (2) confirm that there are at least 10 vowel segments at higher SPL, (3) find the first segment above 80 dB SPL whose SER is within 0.5 dB of the maximum, and (4) select its SPL as the saturation SPL.

Some more specific comments to Figs. 5(a)–5(d) can also be made. (a) Male subject, vowel group {i|y|e}: note how the level increments above 2 kHz reduce to about 3 dB at high effort. This means that the spectrum *balance* SB stops increasing at high effort. (b) Male subject, vowel group {u|o}: note how at high effort the spectrum *level* around 3 kHz stops increasing altogether. The segments with the highest SPL's increase in the low band only. This was observed in about a third of all cases. (c) Female subject, vowel group {a|ɑ}: judging from the bimodal histogram, the subject appears to have clustered around a low and a high SPL, which may account in part for the abrupt jump in the high spectrum. (d) Male, vowel group {a|ɑ}: the high spectrum level increments do not quite reduce to 3 dB, but the F1 peak still becomes more pronounced. This is the same subject as in Fig. 10(b), solid line.

B. Spectrum balance and saturation

The SER saturation-detection algorithm was applied to three vowel groups and 23 subjects, for a total of 69 observations. Saturation was observed in all but one of these. The sound level at which the SER saturated, SPL_{sat}, in the [a] vowel group was 97.4 dB (mean) ±4.9 dB (standard deviation) for males and 93.2±4.0 dB for females (at 0.30 m microphone distance). The same SPL_{sat} averages were 4–5 dB lower for the [i] vowel group, and 2–4 dB lower for the [u] vowel group (Tables II and III). The differences between the vowel groups were consistent with the acoustic theory of the vocal tract [see the caption for Fig. 7(b)].

The SPL_{sat} averaged across vowel groups was 90.3 dB for females and 95.5 dB for males. This gender difference of 5.2 dB is similar to a 4.2 dB gender difference in the average

TABLE II. Spectrum balance statistics of the observed voice spectrum balance at "saturation" SB_{sat}, and of the SPL at which SB_{sat} was reached, SPL_{sat}. The high- and low-frequency bands were separated using third-order bandpass filters. The vowel segments were sorted by increasing SPL and their SB values were smoothed by a 5 dB wide median filter. The point of saturation was computed with the automatic Sorted Energy Ratio procedure (see text).

[dB]	[a] vowels mean (SD)	[i] vowels mean (SD)	[u] vowels mean (SD)
SB_{sat}, females (N=12)	−7.4(5.1)	−1.2(6.7)	−17.6(4.3)
SB_{sat}, males (N=11)	−11.2(2.7)	−5.2(4.1)	−17.0(2.4)
SPL_{sat}, females	93.2 (4.0)	88.3 (4.6)	89.4 (4.2)
SPL_{sat}, males	97.4 (4.9)	93.0 (3.6)	96.1 (5.4)

SPL of the phonetograms that was found in the companion study. In other words, the male voice SPL's were on 4.2 dB higher than for the females, when averaged over all conditions. Both these findings are consistent with the notion that female voices will have more problems than male voices in a given noisy environment.

One may now ask whether spectral saturation can be taken as a personal indicator of overloaded voice. Since the strongest segment in a vowel group might be an outlier point, the 97% high percentile in SPL of all segments in a vowel group, $SPL|_{97\%}$, was arbitrarily selected as a fairly robust (though conservative) estimate of each person's maximum voice SPL. The mean intrasubject difference between $SPL|_{97\%}$ and the same person's SPL_{sat} was 5.7±3.0 dB for females and 5.5±2.5 dB for males when averaged over the three vowel groups. In other words, we found that the spectrum balance saturated at an SPL that was some 6 dB below the SPL of that person's loudest segments (Fig. 9). The standard deviation in this "overload margin" was smaller than the intersubject standard deviation in the SPL_{sat} as such, suggesting that SPL_{sat} can be taken as a predictor of maximum SPL. If each subject's single loudest segment had been used instead of $SPL|_{97\%}$, the "overload margin" would have been 8–9 dB rather than 6 dB.

The average of SB_{sat} as measured with the bandpass filters in the [a] vowel group was −11.2±2.7 dB for males and −7.4±5.1 dB for females (Table II). Females thus had a somewhat higher maximum SB_{sat} than males (a higher proportion of high-frequency energy at saturation). This was probably due to the higher frequency of the F1 peak in the [a] vowels of females that may have caused the F1 peak to "spill over" into the high band, more than for males. On [u] vowels, there was no gender difference in SB_{sat}.

TABLE III. SPL_{sat} statistics, alternative method. As in Table II, these are the statistics of the observed voice SPL at which SB_{sat} was reached, SPL_{sat}, but measured in an entirely different way. Here the spectrum balance was computed using only a first-order high-frequency emphasis filter for the high-band; the SB data were smoothed using polynomials rather than the Sorted Energy Ratio; and the saturation point was determined by manual inspection rather than by an algorithm. Note that the estimates of SPL_{sat} are the same as in Table II to within about 1 dB. This analysis was performed *before* that shown in Table II.

[dB]	[a] vowels mean (SD)	[i] vowels mean (SD)	[u] vowels mean (SD)
SPL_{sat}, females (N=12)	93.0 (3.2)	87.3 (3.7)	91.1 (4.2)
SPL_{sat}, males (N=11)	98.5 (4.1)	93.5 (3.3)	95.9 (3.8)

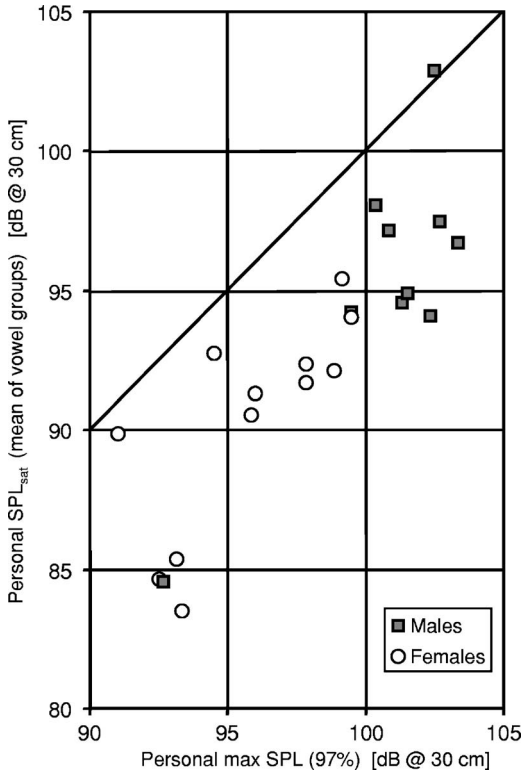


FIG. 9. Comparison of the maximum voice level SPL_{max} (defined as the 97% percentile of all segment SPL's) to the saturation SPL, by subject. Each symbol represents one subject, averaged over the three vowel groups. Note how males (squares) and females (circles) hardly overlap in maximum SPL, and how the SPL_{sat} is 5–6 dB lower than SPL_{max} in most cases. The mean of this "headroom" was 5.5±2.5 dB for males and 5.7±3.0 dB for females.

C. Shape of SB versus SPL

The main issue of this study was to test for the presence of spectrum saturation in loud running speech. Visual inspection of the plots of SB_{median} versus SPL (cf. Fig. 7) confirmed the positive result of the SER-based algorithm. In many cases, SB_{median} would in fact recede again toward the highest SPL's, due to a steep increase of energy at *low* frequencies.

However, these plots were interesting also in other ways. For example, the intra-individual variation across vowel groups was smaller than the inter-individual variation. To facilitate visual comparison, Fig. 10 shows these plots for 4 of the 23 subjects, while similar plots of all 23 subjects are provided at the end of the paper, in Figs. 13 and 14. Although a few subjects exhibited an essentially linear progression of SB_{median} with SPL, this was not the typical case. For a few, SB was practically bimodal, switching from a low to a high value over a small region of voice SPL [Fig. 10(c)]. This may have been due to the subject clustering around either of two voice SPL's ("normal" and "loud"?), regardless of the precise level of the ambient noise. [This is the same subject as in Fig. 5(c). The histograms for the vowel groups [i] and [u] in this subject were not clearly bimodal.] Such clustering would explain the vertical step, but not the horizontal portions in the SB_{median} graph.

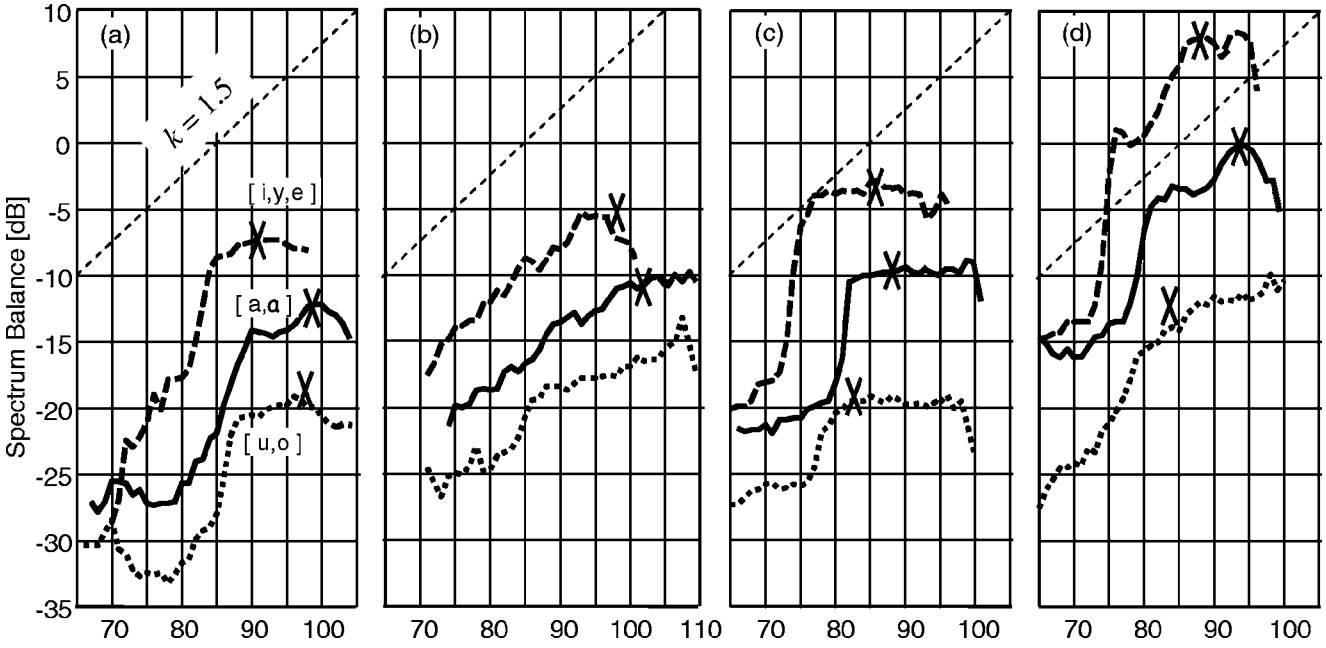


FIG. 10. Median-filtered spectrum balance graphs (see Fig. 7) for four subjects, with vowel groups acting as replications within subjects. The horizontal scale is dB SPL at 30 cm. [i]-like vowels: dashed line; [a]-like vowels: solid line; [u]-like vowels: dotted line. The crosses mark the saturation SPL as determined by the algorithm described in Sec. III C. The diagonal dashed lines indicate $k=1.5$, as discussed in Sec. V D. (a) Male subject M9 with a typical SB progression, containing a steeper portion at medium SPL and a saturating SB toward high SPL. (b) Male subject M3 with less typical, almost linear SB-SPL curves. (c) Female subject F1, with a practically bimodal SB and an abrupt transition from low to high SB. (d) Subject F9, one of two females with particularly high (even positive) SB values at high voice effort. Two to three subjects in both the male and female groups had SB curves of type (b) or (c).

It should be noted that a maximum in the SB_{median} graph need not be due to *source* spectrum saturation: it could also be the combined effect of a gradually peaking F1 in the low band that is "overtaking" an increasing spectrum slope in the high band. For the same reason, a horizontal portion in the SB_{median} graph does not necessarily imply that the phonatory and articulatory configurations remained unchanged with SPL. Explaining the shape of SB-vs-SPL graphs in detail will require further study. The main message of Figs. 10, 13, and 14 is that different individuals exhibited quite different patterns of spectrum balance change with SPL. For instance, there was one "outlier" subject, a female non-native speaker of Swedish, with a very unstable and problematic voice, who had SB_{median} curves that were quite unlike the others, being essentially horizontal for all three vowel groups, and with a restricted range in SPL (subject F8 in Fig. 13).

V. DISCUSSION

A. Perceptual evaluation, self-ratings, and phonetogram results

The results of the perceptual evaluations, self-ratings, and phonetographic analysis have been reported in a companion paper.[18] For the perceptual evaluations, three expert listeners rated the productions of two selected sentences for the perceptual entities strain, press, instability, roughness, and vocal fry. After each reading, the subjects rated their self-perceived success in making oneself heard over the noise, the vocal effort, and vocal strain or fatigue. The phonetographic analysis gave statistics of SPL, F0, and phonation times. Gender differences were found in all the acoustic measures and in some of the perceptual ratings and self-ratings, generally reflecting stronger voices in males.

Interestingly, and perhaps somewhat disappointingly, we could find very little covariation between those results and the spectrum-related results reported here. For example, the SB at saturation did not co-vary with self-rated strain, nor with "making oneself heard." Nor was there any covariation between the saturation results and the perceptual evaluations by experts of the subjects' voices, other than for the "outlier" subject mentioned above.

Only one nontrivial connection was found between results from this study of loud voice spectra and from the companion study. For the males, each subject's subjective self-rating of vocal strain was better predicted by the "overload margin" (voice $SPL-SPL_{sat}$) than by average voice SPL alone, see Fig. 11. This is a potentially interesting finding that is consistent with the notion that SPL_{sat} could be related to a personal strain limit. For the females, the self-rated vocal strain versus SPL was very similar across individuals (Fig. 12), with a majority reporting a sharp increase in vocal strain in the SPL interval 85–90 dB; and in this case normalization to SPL_{sat} did not reduce the variance.

B. Mechanisms for low-frequency emphasis at high effort

At low effort, the vowel spectrum tends to be entirely dominated by the fundamental partial, while at high effort (especially on the [i]-like vowels) the spectrum can be dominated by the F3 region. In this investigation we have found that the low end of the spectrum again becomes more dominant at very high effort levels. We have not found reports of

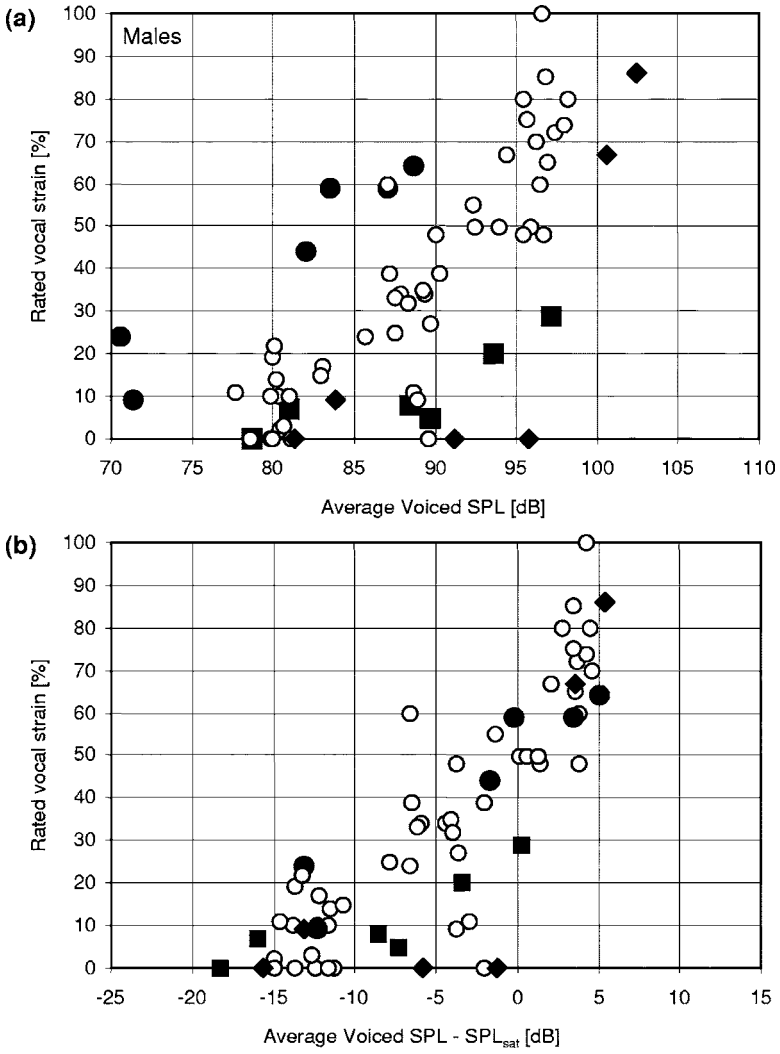


FIG. 11. Subjective self-ratings of vocal strain (males), made after each 90-s noise condition. In (a) the subjective strain is plotted against the SPL at which the subject had spoken in that noise condition (specifically, the time-averaged SPL obtained from phonetograms, of all voiced sounds only). In (b) the same points have been normalized to the individual SPL_{sat}. One subject with low voice SPL is shown with filled circles, and two with high voice SPL are shown with filled squares and diamonds. The remaining eight subjects are plotted with open circles. Note in (b) how normalizing to SPL_{sat} improved the fit to the rated vocal strain.

this elsewhere. A similarly sharpened first formant can, however, be seen in some LTAS of recordings of Croatian folk singing in the *dozivački* style,[22] which is reported to be close to shouting. There are several mechanisms that could raise the low end of the voice spectrum at high effort levels:

1. Increased formant Q values

Higher-than-normal effective Q values of the formants could result from a larger closed quotient at high effort levels. A short open phase implies lower tracheal losses from the formant resonances. However, the closed quotient has

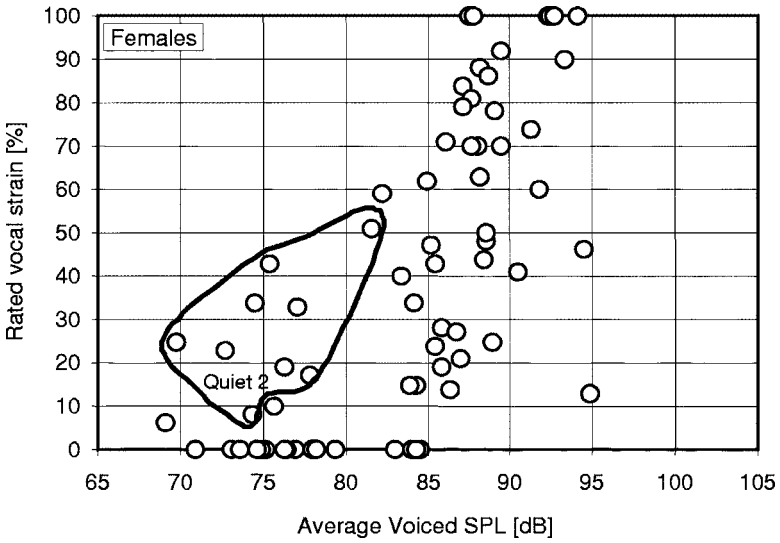


FIG. 12. Subjective self-ratings of vocal strain (females). The subjective strain is plotted against the average SPL at which the subject had spoken in the corresponding noise condition (as in Fig. 11). Points within the closed curve are from the last condition (Quiet 2), without environmental noise, and exhibit some residual voice strain. Note how the strain was reported by most females to rise abruptly at 85–90 dB SPL.

been observed to approach its maximum at voice levels well below the maximum voice level,[23] and so probably does not contribute anything extra at very high efforts.

2. Increased closed quotient

A larger closed quotient means in itself that the amplitude of the second partial relative to that of the first (H2-H1) increases. In cases in which F1 equals about twice F0, this could make H2 unusually high, which is often seen in our recordings. This effect is sometimes referred to as a "glottal formant" (although we would discourage the use of that term), meaning that the source spectrum of theoretical time-domain models such as the LF model can exhibit a hump that rises in frequency with increasing closed quotient, further emphasizing the second, third, or fourth partials. According to Henrich,[24] this emphasis will peak at $2F_0$ for a closed quotient of 0.6, for example.

3. Formant-to-partial matching

At high F_0, whether it be induced by high effort or adopted as a strategy for loudness, the overall voice SPL will begin to depend a great deal on how partials 1–4 happen to coincide with F_1. This was expressed by Titze and Sundberg[10] as a ripple effect in the SPL as a function of F_0. Since the SPL will increase quite markedly when F_1 is close to nF_0, it is conceivable that some speakers could tend to adopt such matching as a strategy for loud voice, for example by preferring a certain range of F_0. This need not even be an active adjustment: in very loud voice, F0 tends to be constrained to a high value (the "top of one's voice"). At the same time, F1 in all vowels tends to be scaled up toward that for [a] because of a larger jaw opening.[3] Hence, if a person's maximum usable F0 happens to line up a partial with F1 in shouting, a habitual match may occur fortuitously. If so, then partials matching F_1 may cause peaking that systematically affects the spectrum balance. Alku *et al.*[25] looked for evidence of formant-to-partial matching, but did not find any. In our recordings, the very loudest vowel segments did in fact often exhibit extremely emphasized second or third partial tones, with the adjacent partials on *both* sides being 20 dB lower or more. However, this could be an artefact of our segmenting algorithm, which selected for analysis the strongest 50 ms of each vowel segment. In cases where the F_0 was changing during the vowel, this would tend to select precisely those 50 ms where a partial happens to match a formant. The influence of this mechanism will require further study. In any case, it seems that the SPL will become a less accurate estimator of vocal effort at high effort than at normal effort.

These mechanisms all tend to emphasize low partials. They also make the study of high effort voice acoustics rather intricate. To isolate the behavior of the source spectrum slope from the articulatory effects, one would probably have to build a detailed mathematical model that includes that slope, together with the above mechanisms, and try to match its output to the observed data, as suggested, e.g., by Traunmüller and Eriksson.[26] This is beyond the scope of the present paper.

C. Relating SB to subglottal pressure

The voice SPL is not the ideal estimator of voice effort. It would be much better to have access to the subglottal pressure p_s during voicing, but this is difficult to combine with a realistic speaking task, and was not attempted for this study. Vilkman *et al.*[2] studied various phonatory variables, including p_s and NAQ, in /pa:p:a/ syllable strings produced by five females and six males over a 40–105 dB range in SPL. NAQ is the peak glottal flow divided by the magnitude of the peak in the first derivative of flow, normalized to F0; it is equivalent to the R_d parameter given earlier by Fant.[27] A low NAQ value is typically associated with more pressed voice and a relative increase in high-frequency energy. The results of Vilkman *et al.* show that NAQ decreased toward loud phonation, corresponding to more rapid closure, but only at low efforts, up to about 80 dB SPL (at 40 cm). At higher SPL, the NAQ did not change much.

The data of Vilkman *et al.* also included the subglottal pressure p_s, estimated from the intra-oral pressure during p-occlusions. On a linear scale, p_s started increasing abruptly above 10 cm H_2O at about 90 dB SPL (at 40 cm), for both females and males. (Some of the figures in Vilkman *et al.* do not carry scale markings, but Vilkman has kindly made the data available to us.) This level is very close to the average SPL_{sat} of 93 dB (at 30 cm) observed for [a] vowels in the present study for females, but less close for male subjects. These data on NAQ and p_s suggest, but do not prove, that at high effort, the glottal source contributes to SPL only by increased flow amplitude, which does not change the spectrum.

The results of Vilkman *et al.*, in combination with ours indicate that the cause for the rising low end of the spectrum at high effort levels in speech would not be phonatory (of the source) but rather articulatory (of the filter). Closer study will be needed to resolve this issue.

D. Comparison to LTAS-based measures of spectrum tilt

Several others[28–31] have found that, in long-term average spectra of running speech or singing obtained from moderate to loud voice, there is an almost linear relationship between the overall level and the level in any narrow part of the spectrum. Specifically, a change of Δx decibels in the total equivalent level has been observed typically to result in a LTAS level change $\Delta y = k(f)\Delta x$ in a band centered on frequency f. The frequency-dependent coefficient k has been reported to range from about 0.5 for frequencies well below the first formant, through 1 at the first formant peak, which usually "carries" the total level, to between 1.5 and 2 for frequencies around 3 kHz; k is usually monotonic but not linear with frequency, up to about 4 kHz.

The SB measure can be related to k for our high-frequency band, 2–6 kHz. Because SB is a level *difference*, the slope of a graph of SB versus SPL will be $k-1$. When $k=1$, SB remains constant with SPL, and the graph of SB versus SPL is horizontal and flat. In Figs. 10, 13, and 14 the aspect ratio has been chosen such that $k=1.5$, which is a typical value reported by others in this frequency band,

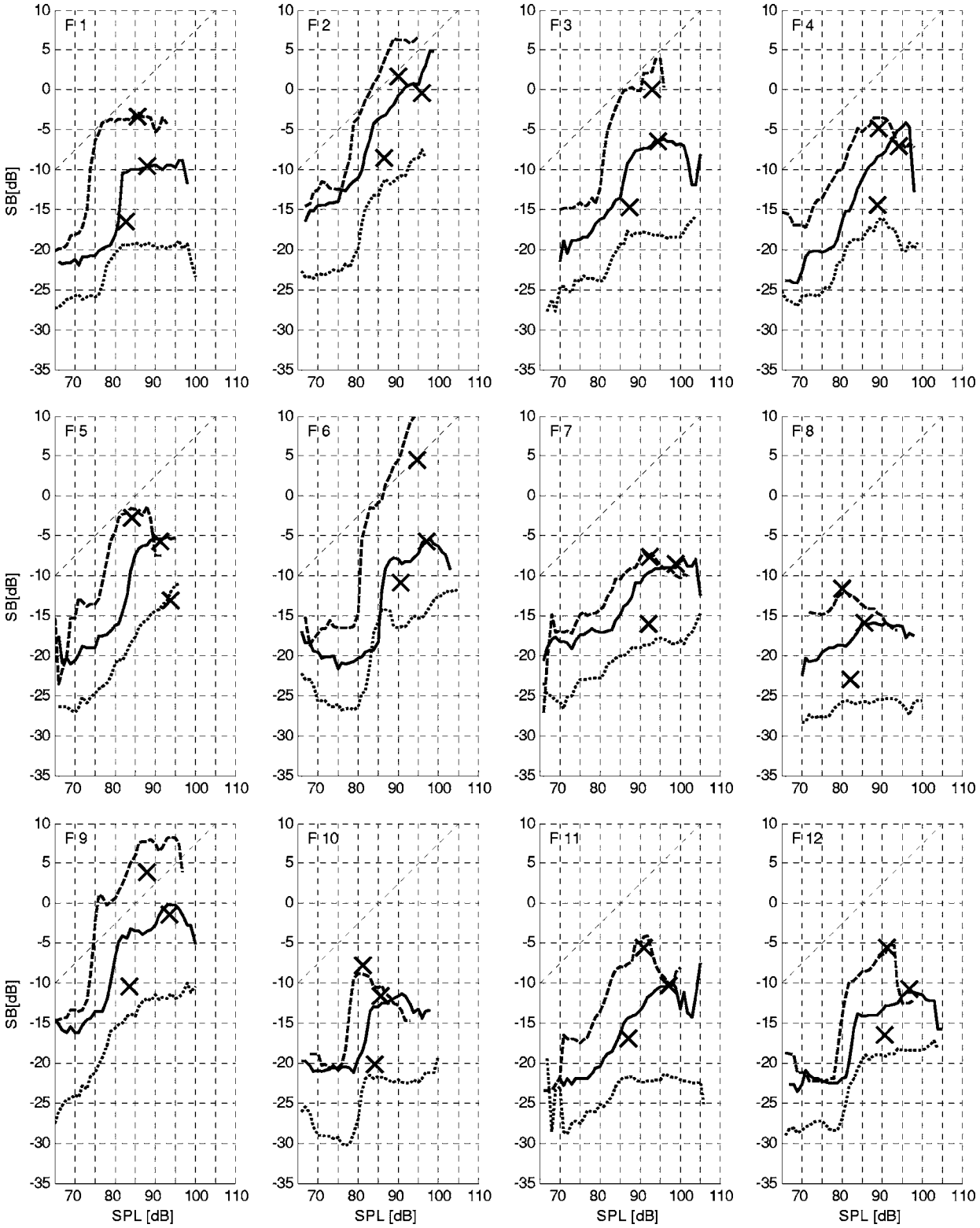


FIG. 13. Spectrum balance (SB) graphs for all female subjects F1–F12, with vowel groups acting as replications within subjects, in the same format as in Fig. 10. The horizontal scale is dB SPL at 30 cm. [i]-like vowels: dashed line; [a]-like vowels: solid line; [u]-like vowels: dotted line. The crosses mark the saturation SPL as determined by the algorithm described in Sec. III C. The crosses do not always fall on the SB curves, because their locations were determined from the sorted energy ratio rather than the spectrum balance. The diagonal dashed lines indicate $k=1.5$, as discussed in Sec. V D. Some of the curves are noisy at the lowest and/or the highest SPL's. The reason is that the SB median filter, which spans across 5 dB in SPL, will contain only a few scattered points at the far ends. Subject F6 on [i]-like vowels saturated at a spectral balance value of +12 dB, just outside the graph.

would correspond to a straight line at a +45° slope. It is clear from the figures that such a straight line is only a first approximation to how SB changed with SPL. More detail was not seen in the LTAS studies, mainly for the reason that the spectra of all vowels were pooled in taking the long *time* average spectrum. From Fig. 7 it can be concluded that, unless vowels from running speech are separated at least into major groups, details of the spectral progression with SPL

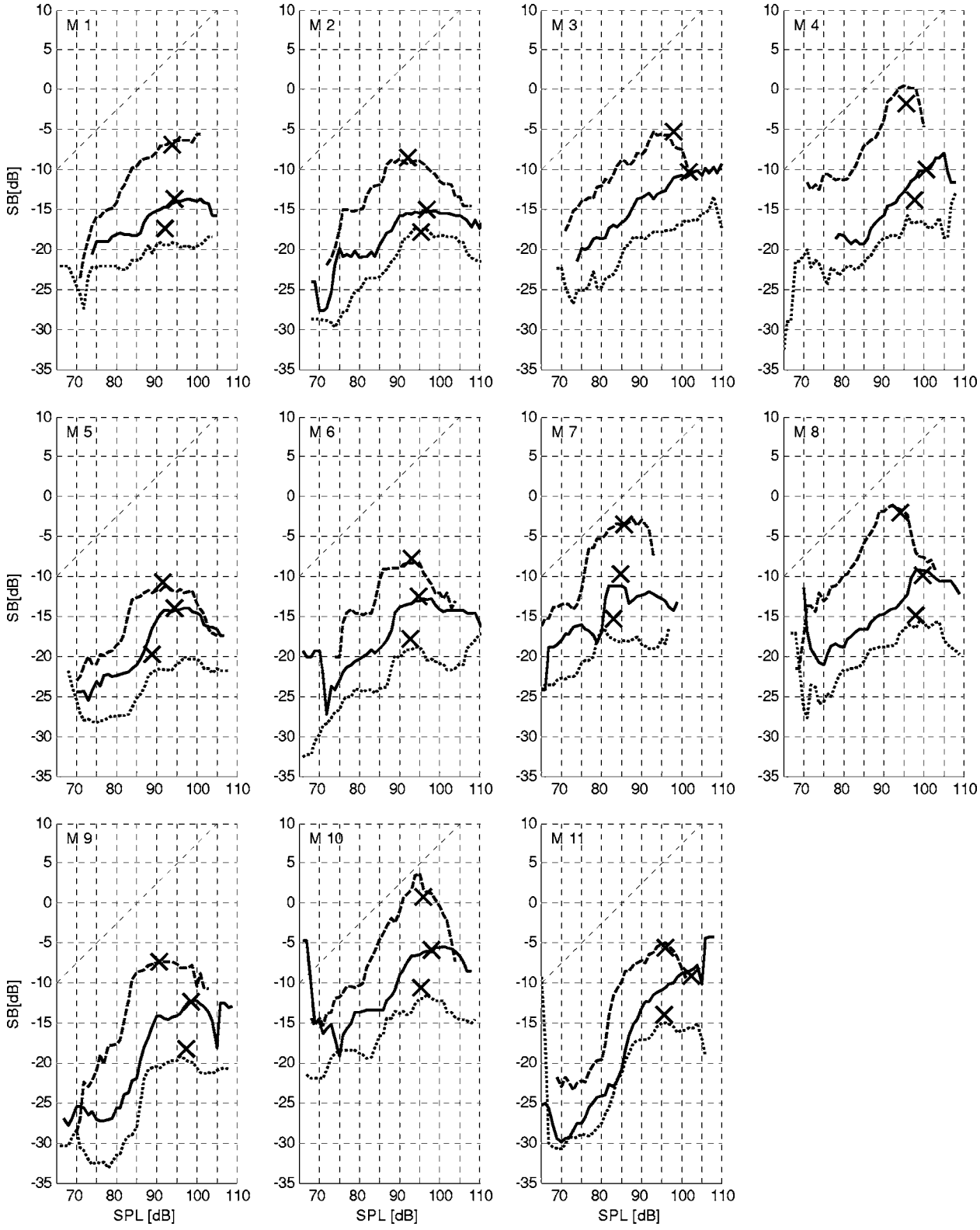


FIG. 14. Spectrum balance curves as in Fig. 13, for all male subjects M1–M11. An absent cross indicates that spectrum saturation was not detected.

will be lost. Other reasons for the lack of detail may include (a) that the LTAS, being an average power measure, tends to reflect the strongest portions of the acquired signal, so low effort levels would not be well represented; (b) that the number of levels sampled was much smaller; and (c) that very high effort levels as in shouting were typically not reached in those studies.

E. On the definition of spectrum balance

Our somewhat *ad hoc* definition of spectrum balance hinges on the characteristics of the band filters. For example, the filters chosen here may incur a bias toward higher (less negative) spectrum balance values with increasing F_0 or F1. This will need further study. However, the shape of the filter

bands is not critical for determining the SPL at which spectrum saturation occurs, as was evidenced by the alternative calculations.

A slightly simpler definition of spectrum balance would be the level difference between L_{high} and the total voice SPL. However, such a measure would never become positive, even if L_{low} were lower than L_{high}, as was found in some very bright voices. This is why we chose to define SB using the levels in two complementary bands. The SB measure is also very similar to some other measures, such as the alpha ratio, which is defined[32] as the ratio of energy above 1 kHz to energy below 1 kHz. For the present study, 1 kHz was thought to be too low a delimiting frequency, since the dominant spectrum peak at high voice effort would sometimes spread above 1 kHz, giving a false value in the high band. By setting the lower limit of the high band to 2 kHz, we hoped to increase the contrast between the bands, but this was only partly effective. It could be argued that a somewhat higher delimiting frequency would be appropriate for female voices.

Traunmüller and Eriksson[3] studied six men and six women speaking or shouting outdoors over distances of up to 180 m. They measured "spectral emphasis" as the difference in dB between total SPL and the SPL of the fundamental alone. This measure did not show any tendency to saturate for voice SPL's up to the maximum reported, nor would it, if the second or third partial became dominant. Unfortunately, their SPL scale was not calibrated.

Yet another possible measure of spectral bias is the spectrum centroid, but the centroid tends to "stick" to a point just above the dominant partial or formant peak, and therefore it may introduce a possibly undesirable covariation with F0 and/or formant frequencies. In some of the vowel groups in this experiment, the spectrum centroid was actually a fairly good estimator of the first formant frequency.

F. Other applications of the noise cancellation technique

The noise cancellation technique can be used in several other ways. Any authentic workplace noise can be recorded separately and used as site noise. Voice patients can be recorded in the noise of the own workplace and the decontaminated recordings can be played to them to make them more aware of what they actually do with their voices. The technique can be combined with speech intelligibility measurements to develop vocal strategies for coping with particular types of noise. Voice therapists who have listened to the decontaminated recordings have been struck by the large degree of voice change that is evident in most subjects. Several have commented on how a voice usage that sounds rather appropriate in a noise will sound quite abusive when the noise is removed. Hence the technique could be of instructional value in voice care.

VI. SUMMARY

Environmental noise, presented realistically over loudspeakers, was used to invoke speech over a large range of effort levels from low or moderate to very high. Running speech, decontaminated from the noise, was obtained from 12 females and 11 males in quiet, in four different conditions of increasing ambient noise, and finally once more in quiet.

In a given noise, males produced 3–5 dB higher SPL than females, and also rated their success in making themselves heard higher than did females. The third-octave-band voice spectrum in one-third of the subjects tended to reach an absolute maximum level in the F3–F5 range well before the maximum overall SPL was reached. At the highest effort levels, an accentuated first formant, possibly in combination with source spectrum factors, resulted in a strong emphasis to one of the lowest few partials.

The "spectrum balance" SB was defined as the level difference between the 2–6 kHz band and the 100–1000 Hz band, with filter skirts of 18 dB/octave. This measure was used as a simpler substitute for spectrum slope. When averaged over many instances of similar vowels, the spectrum balance increased with SPL, but only up to about −7 dB (females) or −11 dB (males) in the [a] vowel group. This maximum occurred at a voice SPL that was quite individual, but on the average at 93.2 dB for females and 97.4 dB for males, again on [a]-like vowels. Above this "spectrum balance saturation level," SPL_{sat}, the spectrum balance stopped increasing, and would sometimes even decrease, especially on [i]-like vowels. The latter phenomenon was due to one or more mechanisms that reinforce the *lower* partials at high effort levels. We did not attempt here to identify any single mechanism responsible for the "spectrum saturation" effect.

For each subject, the self-rated vocal strain in each noise type was of course related to the average voice SPL in that noise type. For males, however, rated vocal strain was better predicted by the SPL normalized to the individual SPL_{sat} than by the SPL alone. For females, such normalization did not result in a better prediction of self-rated strain, apparently because the variation within the group was small.

The spectral data obtained here strongly suggest a marked change of articulatory regime at the highest effort levels, presumably into shouting rather than speaking. Given that each 3-dB increase in voice SPL entails a doubling of the radiated acoustic power, this is hardly surprising. Ideally, we would like to be able to determine some kind of vocal hazard strain limit for each individual. The personal SPL_{sat} measure introduced here seems to have some predictive value in this regard, but it is not straightforward to measure. It would be desirable to device a simpler procedure for detecting spectral saturation in a natural environment. A modified speech recognition system could probably be applied to this purpose.

The important task remains to explain the spectral evolution in more detail, relating it to phonatory mechanisms and their possible operating limits for sustainable loud voice. Noise abatement in the workplace would need to take such a limit into account to be effective for voice purposes. Ultimately, it would be useful to arrive at a model for predicting the amount of vocal strain that a particular workplace noise will engender in a given person. Even a small reduction in workplace noise could make a large difference to the individual worker.

The term "spectrum [balance] saturation" may need to be changed to something more accurate, should the effect be confirmed and explained.

ACKNOWLEDGMENTS

Our thanks go to Professor Johan Sundberg and Professor Gunnar Fant and to Dr. Svante Granqvist, Dr. Jan Švec, and Dr. Nathalie Henrich for valuable comments and discussions, as well as to many other colleagues. Carina Aronsson did the painstaking editing of the text transcriptions, and Kåre Sjölander enthusiastically breathed new life into the automatic speech segmentation program. This work was supported by the Swedish Council for Working Life and Social Research, under Contracts No. 2001-0341 and No. 2002-0416.

[1]N. Roy, R. Merill, S. Thibeault, R. Parsa, S. Gray, and E. Smith, "Prevalence of voice disorders in teachers and general population," J. Speech Lang. Hear. Res. **47**, 281–293 (2004).
[2]E. Vilkman, J. Vintturi, and P. Alku, "Dynamic extremes of voice in the light of time domain parameters extracted from the amplitude features of glottal flow and its derivative," Folia Phoniatr Logop **53**, 144–157 (2002).
[3]H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," J. Acoust. Soc. Am. **107**, 3438–3451 (2000).
[4]R. Schulman, "Articulatory dynamics of loud and normal speech," J. Acoust. Soc. Am. **85**, 295–312 (1989).
[5]B. Frøkjær-Jensen, "Changes in formant frequencies and formant levels at high voice effort," ARIPUC 1 (Inst. Phonet., University of Copenhagen, 1966), pp. 47–55.
[6]J.-S. Liénard and M.-G. di Benedetto, "Effect of vocal effort on spectral properties of vowels," J. Acoust. Soc. Am. **106**, 411–422 (1999).
[7]D. Rostolland, "Acoustic features of shouted voice," Acustica **50**, 118–125 (1982).
[8]D. Rostolland, "Phonetic structure of shouted voice," Acustica **51**, 80–89 (1982).
[9]J. E. Huber, E. T. Statopoulos, G. M. Curione, T. A. Ash, and H. K. Johnson, "Formants of children, women, and men: The effect of vocal intensity variation," J. Acoust. Soc. Am. **106**, 1532–1542 (1999).
[10]I. R. Titze and J. Sundberg, "Vocal intensity in speakers and singers," J. Acoust. Soc. Am. **91**, 2936–2946 (1992).
[11]H. Wilkens and H.-H. Bartel, "Wiedererkennbarkeit der originallautstärke eines sprechers bei elektroakustischer wiedergabe" ("Recognizability of the original loudness of a speaker on electroacoustic reproduction"), Acustica **37**, 45–49 (1977).
[12]G. Fant, *Acoustic Theory of Speech Production* (Mouton, The Hague, 1960).
[13]G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," Speech Transmission Laboratory Quarterly Progress and Status Report, STL-QPSR 4/1985 (1985), pp. 1–13.
[14]G. Fant and Q. Lin, "Frequency domain interpretation and derivation of glottal flow parameters," Speech Transmission Laboratory Quarterly Progress and Status Report, STL-QPSR 2–3/1988 (1988), pp. 1–21.
[15]S. Ternström, M. Södersten, and M. Bohman, "Cancellation of simulated environmental noise as a tool for measuring vocal performance during noise exposure," J. Voice **16**, 195–206 (2002).
[16]B. Hammarberg, *Perceptual and Acoustic Analysis of Dysphonia*, Ph.D. thesis, Karolinska Institute, Stockholm (1986).
[17]M. Bohman, S. Ternström, and M. Södersten, "The use of channel estimation techniques for investigating vocal stress in noisy work environments," *Proceedings of the Joint Baltic-Nordic Acoustical Meeting 2002* (B-NAM 2002), Copenhagen, August (2002). Published in Ultragarsas **3**(48), 9–13, Kaunas University of Technology, Lithuania (2002).
[18]M. Södersten, S. Ternström, and M. Bohman, "Loud speech in realistic environmental noise: phonetogram data, perceptual voice quality, subjective ratings and gender differences in healthy speakers," J. Voice **19**, 29–46 (2005).
[19]K. Sjölander, "Automatic alignment of phonetic segments," Working papers **49**, 140–143, Department of Linguistics, Lund University, Sweden (2001). Available online at http://www.ling.lu.se/disseminations/pdf/49/bidrag36.pdf.
[20]P. Pabon, personal communication. ("Dynamics of spectrum slope in phonetograms," unpublished poster at 4th Pan-European Voice Conference, PEVOC IV, Stockholm, August 2001; available online at http://www.let.uu.nl/~Peter.Pabon/personal).
[21]M. R. Schroeder, "New method of measuring reverberation time," J. Acoust. Soc. Am. **37**, 409–412 (1965).
[22]G. Kovačić, P. Boersma, and H. Domitrović, "Long-term average spectra in professional folk singing voices: A comparison of the *klapa* and *dozvački* styles," *IFA Proceedings* Vol. 25 (Institute of Phonetic Sciences, University of Amsterdam, The Netherlands, 2003).
[23]J. Sundberg, M. Andersson, and C. Hultqvist, "Effects of subglottal variation on professional baritone singers' voice sources," J. Acoust. Soc. Am. **105**, 1965–1971 (1999).
[24]N. Henrich, *Etude de la source glottique en voix parlée et chantée* (Study of the glottal source in speech and singing), Ph.D. thesis in acoustics, Université Paris 6 (2001), pp. 55–82.
[25]P. Alku, J. Vintturi, and E. Vilkman, "On the linearity of the relationship between the sound pressure level and the negative peak amplitude of the differentiated glottal flow in vowel production," Speech Commun. **28**, 269–281 (1999).
[26]H. Traunmüller and A. Eriksson, "A method of measuring formant frequencies at high fundamental frequencies," *Proceedings of EuroSpeech 1997* (1997), Vol. **1**, pp. 477–480.
[27]G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," Speech Transmission Laboratory–Quarterly Progress and Status Report, STL-QPSR 36(2–3), 119–156 (1995). (http://www.speech.kth.se/qpsr/pdf/1995/1995_36_2-3_119-156.pdf).
[28]G. Fant, "The voice source in connected speech," Speech Commun. **22**, 125–139 (1997).
[29]M. Nordenberg and J. Sundberg, "Effect on LTAS of vocal loudness variation," Logopedics Phoniatrics Vocology **29**, 183–191 (2004).
[30]P. White, "The effect of vocal intensity variation on children's voices using long-term average spectrum (LTAS) analysis," Logopedics Phoniatrics Vocology **23**, 111–120 (1998).
[31]S. Ternström, "Long-time average spectrum characteristics of different choirs in different rooms," Voice (UK) **2**, 55–77 (1993).
[32]P. Kitzing, "LTAS criteria pertinent to the measurement of voice quality," J. Phonetics **14**, 477–482 (1986).

Functional data analysis of prosodic effects on articulatory timing

Sungbok Lee
USC Department of Linguistics and USC Viterbi School of Engineering, 3601 Watt Way, GFS 301, Los Angeles, California 90089-1693

Dani Byrd[a] and Jelena Krivokapić
USC Department of Linguistics, 3601 Watt Way, GFS 301, Los Angeles, California 90089-1693

(Received 25 March 2005; revised 7 November 2005; accepted 4 December 2005)

An application of functional data analysis (FDA) (Ramsay and Silverman, 2005, Functional Data Analysis, 2nd ed. (Springer-Verlag, New York)) for linguistic experimentation is explored. The functional time-registration method provided by FDA is shown to offer novel advantages in the investigation of articulatory timing. Traditionally, articulatory studies examining the effects of linguistic variables such as prosody on articulatory timing have relied on comparing the durations of speech intervals of interest defined by kinematic landmarks. Such measurements, however, do not preserve information on the detailed, continuous pattern of articulatory timing that unfolds during these intervals. We present an approach that allows the analysis of entire, continuous kinematic trajectories obtained in a movement tracking experiment examining the influence of a phrasal boundary on articulatory patterning. FDA time deformation functions, after alignment of test and reference (control) signals, reveal delaying of articulator movement (i.e., slowing of the internal clock rate) in the presence of a phrase boundary as the speech stream recedes from the boundary. This is a theoretically predicted pattern (Byrd and Saltzman, 2003, The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening, Journal of Phonetics **31**, 149–180.), which would be more difficult to validate with a traditional interval-based approach. It is concluded that the FDA time alignment method provides a useful tool for characterizing timing patterns in linguistic experimentation based on continuous kinematic trajectories. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2161436]

PACS number(s): 43.70.Jt, 43.70.Bk [AL] Pages: 1666–1671

I. INTRODUCTION

A. Background

In the past, experiments testing for the effects of linguistic variables on the temporal patterning of articulation have relied on comparing the durations of articulatory intervals defined piecewise by kinematic landmarks. For example, a number of articulatory movement tracking studies have shown that lengthening of articulatory movements occur at prosodic boundaries (Edwards *et al.*, 1991; Beckman and Edwards, 1992; Byrd and Saltzman, 1998; Byrd, Kaun, Narayanan, and Saltzman, 2000; Fougeron, 2001; Cho and Keating, 2001; Tabin, 2003; Keating *et al.*, 2004; Cho, in press; and Tabain and Perrier, 2005). The previous articulatory studies of this sort have relied on kinematic landmarks such as movement edges/extrema and peak velocities to define speech intervals of interest and compare their durations. Such measurements, however, lack information on the detailed pattern of articulatory timing that unfolds along the time dimension during the durational periods. It is expected that an examination of articulatory trajectories in a continuous way can reveal such timing evolution and thus can facilitate an understanding of the linguistic patterning.

The statistical framework called functional data analysis (FDA), introduced by Ramsay (1982; followed by Ramsay and Silverman, 1997, Ramsay and Silverman, 2005), offers a novel alternative that can consider entire, continuous kinematic trajectories obtained in various experimental conditions. FDA allows the deformation or warping of these trajectories over time to be characterized and compared within and across subjects in evaluating the linguistic variable of interest. For example, prosodic models that seek to explain how speakers modulate the spatiotemporal organization of articulatory gestures as a function of their phrasal position are particularly informed by examining continuous kinematic trajectories. It has been hypothesized that the internal "clock rate" that controls the temporal unfolding of utterances is slowed as a phrase boundary is approached and speeds up again as the boundary recedes (Byrd and Saltzman, 2003). Because such a change in articulatory dynamics is best described in the continuous time dimension, it is hypothesized that the FDA time alignment method will be able to detect such local temporal fluctuation of gestural activation near prosodic boundaries. In addition, the resulting continuous time warping functions provide data useful in constructing and verifying such models.

While most conventional statistical methods process a

[a]Corresponding author: Dani Byrd, USC Linguistics, 3601 Watt Way, GFS 301, Los Angeles, California 90089-1693. Telephone: (213) 821-1227; electronic mail: dbyrd@usc.edu

 0001-4966/2006/119(3)/1666/6/$22.50

collection of individual data points, the FDA statistical framework is designed to process a collection of functions or curves (Ramsay and Silverman, 2005). The term "functional" reflects a view that by expressing discrete data in a functional form, one can better represent the underlying continuity of the physical or physiological system generating the data. Each curve is regarded as a sample of an underlying *common* pattern. It also permits a more natural way to utilize its derivatives (e.g., velocity and acceleration) for system description or modeling. In practice, such a functional representation of data is achieved by converting the raw sampled data points into a continuous function based on basis function expansion and smoothing.

The FDA framework provides novel data processing and statistical analysis algorithms for the creation and exploration of functional data (Ramsay and Silverman, 2005). Specifically, two essential data processing methods in the FDA framework are functional data smoothing and functional time alignment or time registration methods. These methods have been developed to prepare data for further analysis in the FDA framework, such as functional analysis of variance, functional principal component analysis, and functional canonical correlation analysis. They can be equally useful for other applications in which data smoothing or time registration of sequential data is desirable. In fact, the motivation of this study is to extend the usefulness of the functional time registration method applied for articulatory speech production studies.

The FDA time registration method has been applied in the analysis of lip movements (Ramsay, Munhall, Gracco, and Ostry, 1996), in aligning laryngeal and audio signals (Lucero *et al.*, 1997; Lucero and Koenig, 2000), in the variability analysis of oral airflow data in children's speech (Koenig and Lucero, 2002), and in the variability study of VCV articulation (Lucero and Löfqvist, 2005). However, in these studies the main focus has been either to demonstrate the FDA time registration method or to estimate signal average and variability in an optimal way from repeated productions of the same utterance. Here we present an extended use of the FDA time registration method for the analysis of kinematic articulatory trajectory data obtained in different linguistic conditions. Specifically, we investigate the difference in tongue-tip temporal patterning in two contrasting prosodic environments, namely, in the presence and absence of an intonational phrase boundary.

B. FDA functional data smoothing

Functional data smoothing is the first step of any data analysis in the FDA framework, and its purpose is to convert raw discrete data points into a smoothly varying function. This emphasizes patterns in the data by minimizing short-term deviation due to measurement errors or inherent system noise. We will give a brief mathematical outline of the FDA smoothing method.

In Ramsay and Silverman (1997, 2005), a preferred approach to the functional data smoothing is the classic least square error minimization method augmented with a regularization term or "roughness penalty" for the control of degree of smoothness, and the cost function F to be minimized is set to

$$F(x,y,\lambda) = \sum_j [x_j - y(t_j)]^2 + \lambda \int \left(\frac{d^4}{dt^4} y(t) \right)^2 dt, \qquad (1)$$

where x_j denotes an observed value at time t_j in a discrete data sequence x, $y(t)$ is the function to be estimated from the observed sequence x, λ is a smoothing parameter, and "d^4/dt^4" denotes the fourth-order time derivative. Now the function $y(t)$ is modeled as a linear combination of a set of basis functions,

$$y(t) = \sum_{k=1}^{K} c_k \phi_k(t), \qquad (2)$$

where $\phi_k(t)$ is the kth basis function with weight c_k, and K is the number of basis functions. Then the task of the functional

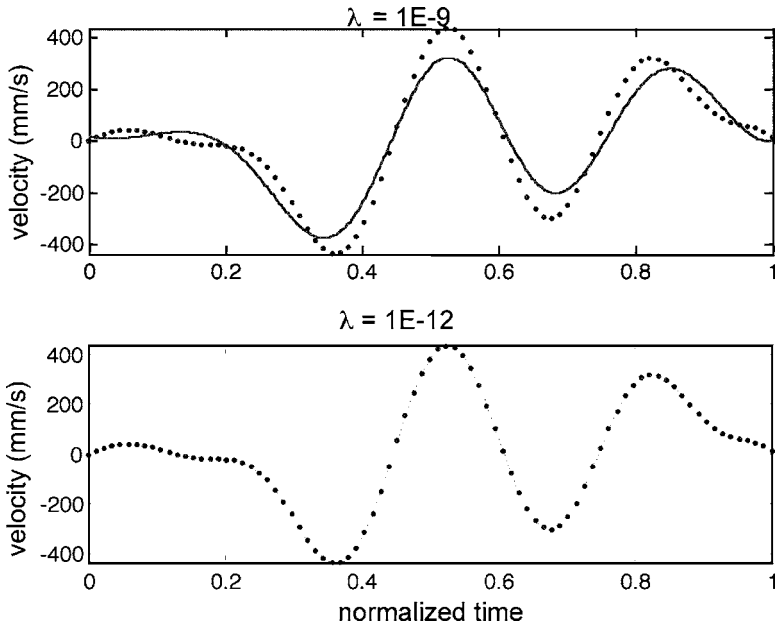


FIG. 1. An example output of FDA smoothing using a tongue tip velocity trajectory. A set of 20 B-splines of order 6 was used for smoothing, but the penalty parameter λ was varied—1E-9 for the top panel and 1E-12 for the bottom panel. The dotted line in each panel represents the original signal and the solid line the smoothed signal. It is illustrated that, for a given number of basis functions and order, the choice of λ is critical to faithfully represent the raw signal.

data smoothing is to find the coefficients c_k, which minimize the cost function F through an iterative minimization procedure.

The choice of basis function depends on the temporal characteristics of data. B-splines (de Boor, 2001) are the typical choice for nonperiodic observation sequences. The smoothing parameter λ is introduced for balance between exact data fitting and smoothing through the roughness penalty.[1] If λ becomes close to zero, a more exact data fitting will occur as the cost function computation is dominated by the minimization of the least square error term. It is noted that the fourth-order time derivative of $y(t)$ is used in the roughness penalty term in order to guarantee the smoothness of the second-order time derivative of $y(t)$, which is related to the curvature of $y(t)$. An output of the FDA data smoothing based on Eq. (1) is illustrated in Fig. 1 for a tongue tip kinematic velocity trajectory. A set of 20 B-spline basis functions of order 6 with two λ values (1E-9 and 1E-12) are tested. The dotted line in each panel represents the original signal. It can be seen that, for a given number of basis functions and order, the choice of λ is critical to faithfully represent the original signal. In fact, the choice of λ was found to be more important in data smoothing than the selection of the order and number of B-splines. By appropriately selecting λ and the order and number of basis functions, one can achieve a flexible approximation of discrete data into a functional form.

C. FDA time registration method

Time alignment or registration refers to an operation by which signals are aligned in time so that a measure of distance between the signals and a reference is minimized. As illustrated in Fig. 2, it is common to observe that signals obtained under the same experimental condition differ in the timings (or phase) and amplitudes of signal landmarks (e.g., major peaks and valleys, zero-crossings), even after duration normalization by an equal-point resampling. The objective of time alignment is to find a common time path between two signals with different properties (one designated the reference signal and one designated the test signal) by expanding or compressing the physical or clock time of a test signal against the reference. The resulting common time path or "time warping function" represents an intersignal timing relation, that is, local advancing or slowing of the internal or system time of a test signal with respect to the physical or clock time of the reference signal. We describe below the conceptual outline of the FDA time registration method adapted for this study. For mathematical details, the reader should refer to Ramsay and Silverman (2005).

Once test and reference signals are represented as functional forms through the FDA smoothing, the task of FDA time registration is to find a smooth time warping function $h(t)$ that minimizes the difference or distance between test and reference signals. In Ramsay and Silverman (1997, 2005), a general approach to FDA time registration is formulated as finding $h(t)$ by minimizing the cost function

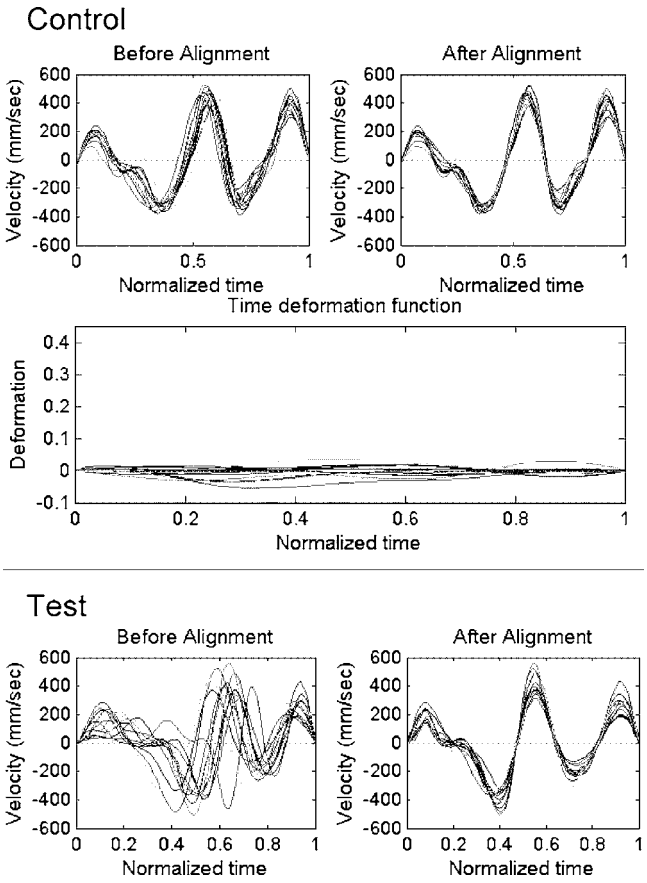


FIG. 2. (Top) Plots of the velocity profiles of control utterances (12 reps) for subject A before time alignment (left) and after time alignment (right). (Bottom) Plots of the velocity profiles of test utterances (12 reps) for subject A before time alignment (left) and after time alignment (right). It is clear that signal average and variability can be measured more accurately after time alignment. (Middle) The middle panel also shows a time deformation function for the control signals compared to reference.

$$D(x,y,\lambda,w) = \int_0^T [x[h(t)] - y(t)]^2\, dt + \lambda \int_0^T w(t)^2\, dt, \quad (3)$$

where $h(t)$ is the time warping function to be determined, λ is a smoothing parameter, $w(t)$ is a smoothness control function for $h(t)$, and T is the end point of the time path. One can note that the form of the cost function is the same as that of the FDA smoothing, that is, the least square minimization augmented with a roughness penalty term.

Since the dimension of $h(t)$ is time, it should be strictly increasing or monotonic and its time derivative should always be positive. Based on these constraints, $h(t)$ can be set to satisfy Eq. (4):

$$\frac{d^2h(t)}{dt^2} = w(t)\frac{dh(t)}{dt}. \quad (4)$$

That is, the first time derivative of $h(t)$, not $h(t)$ itself, is modeled as an exponential growth function, and $w(t)$ controls the behavior of $h(t)$.[2] For instance, when $w(t)$ is positive, the rate of internal time change of the test signal $h(t)$ is slowed when compared to the physical time [i.e., $h(t) > t$], and thus the test signal runs "late." That is, the same landmark occurs later in clock time when compared to the reference signal. It is noted that the square of $w(t)$ is used as the

TABLE I. Sentence 2 is testing the rightward effect: the boundary is after the first consonant, and the consonants to be measure are D D N. Sentence 1 is the control sentence and contains no boundary.

Effect	Consonants	Sentence
Control	NDDN	Birdhunting, we were shocked to see a **n**ew **d**o**d**o **kn**ocking on wooden posts
Rightward effect	N #DDN	At the zoo, we were shocked to see a **Gn**u. **D**o**d**o **kn**ocking about, however, would have been more surprising

regularization term in Eq. (4), which is equivalent to the square of "relative curvature" of $h(t)$ [i.e., the second time derivative of $h(t)$ scaled by its first derivative].

The general solution of Eq. (4) is obtained by integrating it twice; the solution is

$$h(t) = C_0 + C_1 \int_0^T \exp\left[\int_0^u w(v)dv \right] du, \tag{5}$$

where C_0 and C_1 are so determined that $h(0)=0$ and $h(T)=1$. C_0 represents a linear time shift, and T is the end point of normalized time. T can be set to 1 without a loss of generality if durations of test and reference signals are normalized before time registration. This is a usual practice in the FDA time alignment procedure.

Now, the task of finding the monotonic time warping function $h(t)$ is reduced to the task of determining $w(t)$. For that purpose, $w(t)$ is expressed as a linear combination of basis functions as in Eq. (2), and $h(t)$ can be determined from $w(t)$ which minimizes the cost function given in Eq. (3).

Because our major interest is in timing, we focus on timing differences in landmarks occurring in both test and reference articulator velocity patterns. Therefore, the landmark time registration with the aforementioned monotone smoothing method is used in this study in order to take advantage of the clear landmark locations observed in the velocity patterns. The landmark time registration accepts predetermined signal landmark time points as break points, and performs time alignment between two adjacent landmark points by linear shifting and scaling of the basis functions. Twelve B-splines of the order 4 and λ value of 1E-12 are used to represent $w(t)$ in this study. All the computations are based on MATLAB implementations of the FDA smoothing and time registration algorithms publicly available at ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns/.

II. METHOD

A. Speech materials

A subset of speech materials described in a previous study (Byrd *et al.*, 2004, submitted) were used and the stimuli are given in Table I. The goal was to study rightward phrase boundary effects using sentences with the same phonological string varying in the presence or absence of an intonational phrase boundary.

The target sequence in each sentence was [...nV dV dV nV...]. The Carstens Articulograph (AG200) was used to track a sensor adhered to the tongue tip. Sensors were also tracked on the maxilla and bridge of the nose for head movement correction, and a sample of the occlusal plane of each subject was acquired. Sensor position was sampled at 200 Hz during articulation. After data collection, the tongue-tip sensor position data was corrected for head movement and rotated to the occlusal plane. The tongue tip y (vertical) signal was differentiated in order to derive the tongue-tip movement velocity. The position and velocity data were smoothed before and after differentiation with a ninth-order Butterworth filter of cutoff frequency 15 Hz. Four native speakers of American English participated. Subjects read each sentence 12 times and were instructed to read in a casual, conversational style. Subjects will be referred to as Subject A, Subject D (the second author), Subject E, and Subject J.

Because this experiment was designed to investigate tongue tip trajectories for alveolar consonants, we will denote the target sequence for ease of presentation as [D D N]. The control sentence contained the sequence [D D N] with no preceding phrase boundary. To examine the rightward effect of the phrase boundary, the consonants [#D D N] with a preceding intonational phrase boundary (sentence 2) were compared against the same sequence in the no-boundary control utterance. Each target sequence [D D N] (from the onset of /d/ to the closure of /n/) is identified in all sentences. The initial edge was defined as the zero-crossing associated with the peak tongue tip raising movement for [d] and the final edge as the zero-crossing associated with the peak tongue tip raising movement for the [n]. Then the velocity signals were processed for each subject using the FDA time alignment procedure described in the next section.[3]

B. Time alignment procedure

First, a linear time normalization is applied to each individual velocity signal by resampling so that each signal has 200 equally sampled data points (see Fig. 1). Twenty B-splines of the order 6 and λ value of 1E-12 are used for smoothing. A reference signal for each subject is then determined from control signals (those without a phrase boundary before the target string) as follows: Initially, an average of the control signals is computed and used as an initial reference signal for time alignment. After time alignment, an average of time-aligned control signals is computed again and used as a reference. Next, test signals (those having a phrase boundary before the target string) are subject to the landmark time registration with respect to the reference signal, and each time warping function is computed against the refer-

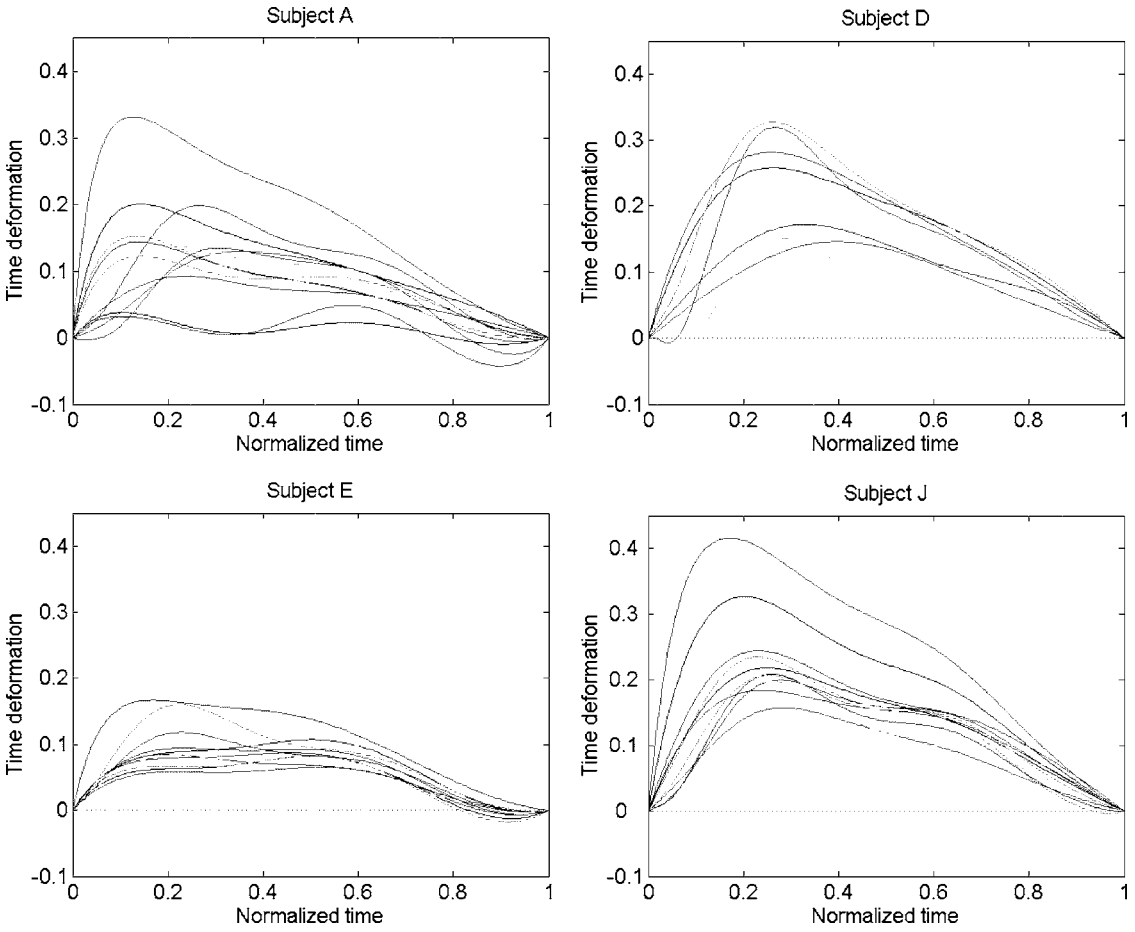


FIG. 3. Time deformations of test signals with respect to the reference for subject A (top left), subject D (top right), subject E (bottom left), and subject J (bottom right). The plots indicate delay or slowing (deformation in the positive direction) of tongue tip articulation in the test condition as compared to the control condition. Note the general asymmetric shapes of deformation patterns. Variability among repetitions is also observed for the degree of slowing, but the slowing shows a similar patterning over time.

ence signal. For landmarks, the internal four zero-crossings of each test signal (see Fig. 2) are used as internal break points. The zero-crossings selected as landmarks correspond to time points where the tongue tip is about to move away from the position extrema. After time alignment, a time deformation function $F_{\text{test}}(t)$ is computed as follows:

$$F_{\text{test}}(t) = h_{\text{test}}(t) - h_{\text{ref}}(t), \tag{6}$$

which represents a delay $[F_{\text{test}}(t) > 0]$ or advance $[F_{\text{test}}(t) < 0]$ of the internal clock time of a test signal with respect to the reference.

It may be noted that there are several ways to do time alignment between two groups of signals: select one typical control signal as a reference and compare it with all test signals, or compare averaged signals (i.e., averaged control and averaged test signals), or, as in this study, compare test signals to a reference signal that is created by averaging the control signals. We chose the last method because the variability in timing among control signals is much less than that between test and control signals or within the set of test signals (see Fig. 2); therefore, the use of an averaged reference signal can be justified. Further, whereas the global alignment procedure (i.e., time alignment without landmarks) was used to create the reference signal, landmark alignment was chosen to compare test to reference due to the high temporal variability present in the test signals, the relative computational efficiency of the landmark alignment procedure, and the relevance of articulatory events to this study. In Fig. 2, tongue tip velocity profiles for control and test utterances are shown before and after FDA time alignment (subject A).

III. RESULTS

In Fig. 3 the resulting time deformation functions of each individual test signal for subjects A, D, E, J are shown. [Figure 2 (middle panel) shows a comparable deformation function for control utterances only (Subject A).] Because a linear time normalization is done before the time alignment, the resulting time warping or deformation function reflects nonlinear, local timing variations in tongue tip closing and release gestures. It is noted that because end points for this analysis are anchored or "pinned" at the edges of the interval of interest, timing effects at the two end points of the overall interval of interest are not discernable.

One can clearly observe detailed patterns of delay relative to the reference pattern of articulator movement as the speech stream recedes from the phrase boundary (recall that the initial and final end points are fixed and are not informative). Generally these temporal modifications due to the pres-

ence of the phrase boundary become reduced as time elapses, i.e., the temporal perturbation is largest close to the boundary and diminishes more remotely (Byrd and Saltzman, 2003) as can be seen by the skew of the time deformation functions and the steeper onset slope than falloff slope of the functions. It is also observed that although there are differences in the amount of time deformation among repetitions, the patterns are fairly similar across repetitions for most subjects. These observations would be difficult to isolate with the conventional landmark-based articulatory timing measurements.

IV. CONCLUDING REMARKS

We conclude from the results of this articulatory kinematic experiment analyzed with FDA time registration that rightward prosodic effects on articulation are greatest locally at the boundary and decrease with distance from the boundary. Such a pattern of delay adjacent to a phrase boundary diminishing with distance from the boundary is predicted within the prosodic π-gesture model of Byrd and Saltzman (2003).

In the past, experiments testing for the effects of linguistic variables on the temporal patterning of articulation have relied on comparing the durations of intervals defined piecewise by kinematic landmarks such as movement edges/extrema and peak velocities. FDA offers a novel alternative that considers entire, continuous kinematic trajectories obtained in various experimental conditions so that the deformation or warping of these trajectories over time can be characterized and compared within subject as well as across subjects.

The FDA smoothing and time registration method can be potentially useful for any applications where data smoothing or time alignment of articulatory trajectories is desirable. In addition, the coefficient set obtained by the FDA smoothing can be used as feature vectors for pattern classification or categorization of time series data including articulatory trajectories. Finally, by integrating the time deformation functions one also can quantify the degree of time deformation, which should be useful to model the strength of boundary types. In the future, we will explore the use of time derivatives of the time deformation function in conjunction with kinematic trajectories for describing articulatory dynamics associated with linguistic conditions.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of NIH Grant No. DC03172, the assistance of Dr. James Mah and Professor Shri Narayanan, and the comments of a very helpful anonymous reviewer.

[1]Because the criteria for an optimal degree of smoothness seems fuzzy and can be subjective and problem dependent, it is unclear whether an automatic way to estimate some essential smoothing parameters (e.g., the degree and number of basis functions and λ) can be optimally formulated within the FDA framework.

[2]In general, an exponential growth function $g(t)$ is modeled as $dg(t)/dt = rg(t)$, where r is a growth rate. One more time derivative of the exponential growth model with a time-dependent growth rate will yield the model equation of the time warping function $h(t)$.

[3]It is noted that the velocity signal has been chosen for time alignment for two main reasons. First, the velocity pattern has traditionally been used for the analysis of skilled movements because underlying dynamic parameters that describe the motion can be derived from the velocity patterns (cf., Nelson, 1983). Second, a velocity pattern has well-defined landmarks (e.g., extrema and zero crossings), which facilitates the FDA landmark time registration. We confirmed that as long as the same landmark time points are used, the time deformation functions estimated from position are effectively the same as those estimated using the velocity pattern.

Beckman, M. E., and Edwards, J. (**1992**). "Intonational categories and the articulatory control of duration," in *Speech Perception, Production and Linguistics Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Ohmsha, Tokyo, Japan), pp. 359–375.

Byrd, D., Kaun, A., Narayanan, S., and Saltzman, E. (**2000**). "Phrasal signatures in articulation," in *Papers in Laboratory Phonology V. Acquisition and the Lexicon* edited by M. B. Broe, and J. B. Pierrehumbert (Cambridge University Press), pp. 70–87.

Byrd, D., Krivokapić, J., and Lee, S. (**submitted**). "How far, how long: On the temporal scope of phrase boundary effects."

Byrd, D., and Saltzman, E. (**1998**). "Intragestural dynamics of multiple phrasal boundaries," J. Phonetics **26**, 173–199.

Byrd, D., and Saltzman, E. (**2003**). "The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening," J. Phonetics **31**, 149–180.

Cho, T. (**in press**). "Manifestation of prosodic structure in articulation: Evidence from lip movement kinematics in English, in *Laboratory Phonology 8: Varieties of Phonological Competence*, edited by L. Goldstein (Walter De Gruyter Inc., New York).

Cho, T., and Keating, P. (**2001**). "'Articulatory and acoustic studies on domain-initial strengthening in Korean," J. Phonetics **29**, 155–190.

de Boor, C. (**2001**). *A Practical Guide to Splines*, revised edition (Springer-Verlag, New York).

Edwards, J., Beckman, M. E., and Fletcher, J. (**1991**). "The articulatory kinematics of final lengthening," J. Acoust. Soc. Am. **89**, 369–382.

Fougeron, C. (**2001**). "Articulatory properties of initial segments in several prosodic constituents in French," J. Phonetics **29**, 109–135.

Keating, P., Cho, T., Fougeron, C., and Hsu, C. (**2004**). "Domain-initial articulatory strengthening in four languages," in *Phonetic Interpretation (Papers in Laboratory Phonology VI)*, edited by J. Local, R. Ogden, and R. Temple (Cambridge University Press, Cambridge), pp. 143–161.

Koenig, L. K., and Lucero, J. L. (**2002**). "The use of functional data analysis to study variability in children's speech: Further data," J. Acoust. Soc. Am. **111**, 2478.

Lucero, J., and Koenig, L. (**2000**). "Time normalization of voices signals using functional data analysis." J. Acoust. Soc. Am. **108**, 1408–1420.

Lucero, J. L., and Löfqvist, A. (**2005**). "Measures of articulatory variability in VCV sequence, ARLO **6**, 80–84.

Lucero, J., Munhall, K., Gracco, V., and Ramsay, J. (**1997**). "On the registration of time and the patterning of speech movement," J. Speech Lang. Hear. Res. **40**, 1111–1117.

Nelson, W. L. (**1983**). "Physical principles for economies of skilled movements," Biol. Cybern. **46**, 135–147.

Ramsay, J. O. (**1982**). "When the data are functions." Psychometrika **47**, 379–396.

Ramsay, J. O., Munhall, K. G., Gracco, V. L., and Ostry, D. J. (**1996**). "Functional data analysis of lip motion," J. Acoust. Soc. Am. **99**, 3718–3727.

Ramsay, J. O., and Silverman, B. W. (**1997**). *Functional Data Analysis* (Springer-Verlag, New York).

Ramsay, J. O., and Silverman, B. W. (**2005**). *Functional Data Analysis*, 2nd ed. (Springer-Verlag, New York).

Sakoe, H., and Chiba, S. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition." IEEE Trans. Acoust., Speech, Signal Process. Vol. **26**, 43–49.

Strik, H., and Boves, L. (**1990**). "A dynamic programming algorithm for time-aligning and averaging physiological signals related to speech," J. Phonetics **19**, 367–378.

Tabain, M. (**2003**). "Effects of prosodic boundary on /aC/ sequences: articulatory results." J. Acoust. Soc. Am. **113**, 2834–2849.

Tabain, M., and Perrier, P. (**2005**). "Articulation and acoustics of /i/ at prosodic boundaries in French," J. Phonetics **33**, 77–100.

West, P. (**1999**). "The extent of coarticulation of English liquids: an acoustic and articulatory study," ICPhS99, San Francisco.

Phonological systems in bilinguals: Age of learning effects on the stop consonant systems of Korean-English bilinguals[a)]

Kyoung-Ho Kang[b)] and Susan G. Guion
Department of Linguistics, University of Oregon, Eugene, Oregon 97403-1290

(Received 29 August 2005; revised 21 December 2005; accepted 21 December 2005)

Interaction of Korean and English stop systems in Korean-English bilinguals as a function of age of acquisition (AOA) of English was investigated. It was hypothesized that early bilinguals (mean AOA=3.8 years) would more likely be native-like in production of English and Korean stops and maintain greater independence between Korean and English stop systems than late bilinguals (mean AOA=21.4 years). Production of Korean and English stops was analyzed in terms of three acoustic-phonetic properties: voice-onset time, amplitude difference between the first two harmonics, and fundamental frequency. Late bilinguals were different from English monolinguals for English voiceless and voiced stops in all three properties. As for Korean stops, late bilinguals were different from Korean monolinguals for fortis stops in voice-onset time. Early bilinguals were not different from the monolinguals of either language. Considering the independence of the two stop systems, late bilinguals seem to have merged English voiceless and Korean aspirated stops and produced English voiced stops with similarities to both Korean fortis and lenis stops, whereas early bilinguals produced five distinct stop types. Thus, the early bilinguals seem to have two independent stop systems, whereas the late bilinguals likely have a merged Korean-English system. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2166607]

PACS number(s): 43.70.Kv, 43.70.Fq [BHS] Pages: 1672–1683

I. INTRODUCTION

Many studies have demonstrated that the two languages of a bilingual may influence each other in speech production. For example, production characteristics of the first language may be inappropriately transferred to the second language or acquisition of a second language may influence the production of the first language. Furthermore, the influence between a bilingual's two languages may increase with age of learning of the second language (L2). There is emerging evidence that earlier bilinguals may have more independent segmental systems for their two languages than later bilinguals. The current study investigates this possibility with an analysis of segmental speech production by Korean-English bilinguals who learned English early ($M=3.8$ years old) or late ($M=21.4$ years old). Specifically, the acoustic properties of Korean and English stops produced by early and late bilinguals are investigated. The early bilinguals are predicted to show less mutual influence between the two stop systems than the late bilinguals. That is, the early bilinguals are predicted to be native-like in both English and Korean due to a greater independence of their two languages. The late bilinguals, on the other hand, are predicted to have non-native-like production of some aspects of both English and Korean due to greater mutual influence of their two languages.

It has been well demonstrated that later learners show more first language (L1) effects on L2 in segmental speech production than early learners. For example, in a study investigating Spanish learners of English, Flege (1991) has shown that early learners' voice-onset time (VOT) values for English /t/ are not distinctive from those of English monolinguals, whereas late learners produce /t/ with values that are intermediate between the short-lag values of Spanish (L1) and the long-lag values of English (L2). Additionally, MacKay *et al.* (2001) examined the production of English /b/ and found that late Italian (L1)-English (L2) bilinguals produced English /b/ with prevoicing more often than the early bilinguals, who, in turn, produced more prevoicing than the English monolinguals. In a study investigating L2 vowel production, Piske *et al.* (2002) examined Italian-English bilinguals' production of 11 English vowels. The production of late bilinguals was significantly less authentic than early bilinguals, whose vowels did not differ from those of English monolinguals. Other studies also have confirmed that early learners are more likely to be native-like in L2 segmental speech production than late learners (e.g., Flege, Mackay, and Meador, 1999; Munro, Flege, and Mackay, 1996, for vowels; Flege, Munro, and MacKay, 1995, for consonants).

In addition to L1 effects on L2, several studies have reported that L2 can affect L1 speech production. MacKay *et al.* (2001) found that both early and late Italian learners of English differed in their Italian stop production from Italian monolinguals. Specifically, they produced Italian /b/ with full prevoicing less often than Italian monolinguals. Several other studies have also found that sequential bilingual children as well as highly proficient late second language learners produce stops in their native language differently from monolingual speakers of that language (Flege, 1987b; Flege and Eefting, 1987a, 1987b; Flege and Hillenbrand, 1984;

[a)]Portions of this work were presented at the Western Conference on Linguistics, University of Southern California, Los Angeles, California, November 2004, and the 149th meeting of the Acoustical Society of America, Vancouver, Canada, May 2005.

[b)]Electronic mail: kkang@uoregon.edu

 0001-4966/2006/119(3)/1672/12/$22.50

Mack, 1990; Williams, 1979; see Guion, 2003 for a review of studies reporting L2 effects on L1 production).

Findings that L2 and L1 can mutually affect each other in speech production have been interpreted as an indication that bilinguals may have a combined L1-L2 phonological system. For example, Harada (2003) reported the effect of an interaction between English and Japanese stop systems in early Japanese (L1)-English (L2) bilinguals. The Japanese-English bilinguals produced Japanese voiceless stops with longer VOT values than Japanese monolinguals but had native-like VOT values for English voiceless stops. These results were interpreted to be an effect of an effort to maintain phonetic contrast between Japanese voiceless stops and English voiced stops in the bilinguals' common phonological space.

Also investigating the proposal that the L1 and L2 may be combined in a single phonological system in bilinguals, Flege, Schirru, and Mackay (2003) examined the interaction of Italian (L1) and English (L2) vowels in bilinguals. Early bilinguals (who used L1 rarely) produced English /eɪ/ with greater formant movement than English native speakers, whereas late bilinguals produced this vowel with less movement. The production with greater formant movement was attributed to an effort to dissimilate the acquired phonetic category for English /eɪ/ from the Italian /e/. The production with less formant movement was attributed to an assimilation of the English /eɪ/ with the Italian /e/.

It also seems that the mutual effects of L1 and L2 on speech production in bilinguals are related to age of learning. There is some evidence that earlier bilinguals may have greater independence between their two phonological systems than later bilinguals. First, later bilinguals may be more likely to use the same phonetic categories in both the L1 and L2, a type of equivalence classification, whereas earlier bilinguals may be more likely to acquire new categories for the L2 that are distinct from the L1 categories (Flege, 1987a, 1987b, 1995).

Second, there is also some evidence that the independence of the acquired L2 phonetic categories from the L1 phonetic categories varies with age. Guion (2003) investigated age of learning effects on the independence of phonological systems by assessing the vowel production ($F1$ and $F2$) of Quichua-Spanish bilinguals. Simultaneous bilinguals were found to have a more finely partitioned vowel space than early bilinguals, such that L1 and L2 front and back vowels were distinct from one another and similar to (near-) monolingual speakers. On the other hand, bilinguals who began learning Spanish in early to late childhood produced Spanish L2 vowels in a native-like way but produced their Quichua (L1) vowels differently from (near-) monolingual speakers. This difference in vowel production was interpreted to be a result of a reorganization which served to enhance perceptual distinctiveness between the vowels of the combined Quichua-Spanish vowel system. Finally, the late bilinguals were found to have merged Spanish and Quichua vowels and seemed to be using their Quichua vowels in Spanish production. These results indicated that simultaneous bilinguals were more likely to have two independent vowel systems than early and late bilinguals. Early and late bilinguals demonstrated interaction between their two vowel systems: the early bilinguals due to interaction between the L1 and L2 systems and the late bilinguals due to the use of the L1 system for L2 production.

However, not all studies have found the two phonological systems of simultaneous bilinguals to be independent. Sundara, Polka, and Baum (in press) investigated the production of coronal stops (/t/ and /d/) in Canadian English and Canadian French by adult simultaneous English-French bilinguals. They found that the bilinguals produced the target stops with language-specific differences between Canadian English and Canadian French but that the bilinguals were not completely identical to monolinguals, indicating possible interaction between the two language systems. Developmental studies also suggest that the phonological systems of simultaneous bilinguals may not develop with full independence from one another (Kehoe, Lleó, and Rakow, 2004; Khattab, 2002).

Thus, although late bilinguals are usually found to have merged phonological systems, the extent of independence between the two phonological systems for early or simultaneous bilinguals is not fully understood. The current study continues investigation into age of learning effects on the interaction of two phonological systems in bilinguals by assessing the interaction of the Korean and English stop systems in early and late Korean-English bilinguals. Specifically, the production of Korean and English stops by early and late Korean (L1)-English (L2) bilinguals is investigated. It is predicted that early bilinguals are more likely to be native-like in their production of English and Korean stop consonants than late bilinguals, and also that early bilinguals have more independent stop systems for their two languages than late bilinguals.

TABLE I. Korean stop consonants.

	Bilabial	Alveolar	Velar
Aspirated	p^h	t^h	k^h
Lenis	p	t	k
Fortis	p^*	t^*	k^*

II. THE STOP CONSONANTS OF KOREAN AND ENGLISH

The consonant system of the Korean language contrasts three types of stops. All three are voiceless in word-initial position, but differ from each other in terms of degree of aspiration and quality of the following vowel. The first type, referred to as *aspirated*, is generally described as strongly aspirated. The second type, referred to as *lenis*, is generally described as slightly aspirated with an onset of breathy phonation in the following vowel. The third type, referred to as *fortis*, is generally described as unaspirated, with an onset of tense phonation in the following vowel. All three stop types occur at three places of articulation: bilabial, alveolar, and velar. Table I presents these stop consonants. The voiceless fortis stops are designated with the diacritic " * " and the voiceless lenis stops are left unmarked, following the con-

ventions used by Cho, Jun, and Ladefoged (2002) and Kim, Beddor, and Horrocks (2002).

Acoustic characteristics of these stops have been well documented in numerous studies. First, regarding voice-onset time (VOT), it has been reported that mean VOT values in word-initial position are shortest for fortis, longer for lenis, and longest for aspirated stops: 15–40, 50–70, and 80–120 ms, respectively (e.g., Cho *et al.*, 2002; Han and Weitzman, 1970; Kim, 1965; 1994; Lisker and Abramson, 1964; Silva, 1992). However, recent work by Silva and colleagues (Silva, 2002; Silva, Choi, and Kim, 2004) suggests that younger Korean speakers may be shortening the VOT for aspirated stops, such that the VOT is similar to lenis stops.

As for the relative difference in amplitude across first harmonic (H1) and higher harmonics (H2 and above) at the onset of a following vowel, lenis and aspirated stops show greater values than fortis stops (Ahn, 1999; Cho *et al.*, 2002; Kagaya, 1974). Generally, the values for lenis and aspirated stops are positive, and negative for fortis stops. This difference indicates that vowels following lenis and aspirated stops are produced with a more breathy quality and vowels following fortis stops are produced with a more creaky quality.

As for *F*0, it is generally agreed that *F*0 at the onset of the following vowel is lowest for lenis, higher for fortis, and highest for aspirated stops (Kim *et al.*, 2002). Ahn (1999), Cho *et al.* (2002), and Shimizu (1996) found that *F*0 for lenis stops was significantly lower than aspirated and fortis stops.

English contrasts two stop categories: voiceless and voiced stops. These categories are reliably differentiated by VOT word initially. For example, Klatt (1975) reported the average VOT of 61 ms for English voiceless stops and 18 ms for voiced stops in stressed monosyllabic words. Lisker and Abramson (1964) also found a great difference between English voiceless (about 80 ms) and voiced (about 15 ms) stops (see also Flege and Eefting, 1986, 1988; Williams, 1977).

As for the H1-H2 measure of the initial portion of the vowels following stops, Ahn (1999) found significantly greater values (a more breathy phonation) for English voiceless stops than English voiced stops. A normalized H1-H2 difference showed a significant difference during the first 1/4 of the total length of the vowel.

As in the case of Korean stops, there are some studies suggesting that *F*0 difference at the onset of the following vowel may play an important role in differentiating English voiceless stops from English voiced stops: *F*0 at the onset of the following vowel is higher for aspirated voiceless stops than for voiced stops (Abramson and Lisker, 1985; Lehiste and Peterson, 1961; Ohde, 1985; Shimizu, 1996).

When investigating age of learning effects in the production of stop consonants, previous studies have mostly considered a single phonetic property, namely VOT (e.g., Flege, 1991; Harada, 2003; Kehoe *et al.*, 2004; Khattab, 2002, 2003; MacKay *et al.*, 2001, with the notable exception of Sundara *et al.*, in press). As described above, Korean stops exploit multiple acoustic cues, such as VOT, voice quality of the following vowel (as indexed by H1-H2 difference), and fundamental frequency of the following vowel (*F*0). Accordingly, the present study measured these various acoustic properties for Korean and English stop consonants.

Also, previous studies have investigated VOT production of bilinguals whose two languages differed in that one of the languages had stops with lead voicing and the other did not (e.g., Khattab, 2002 for English and Arabic, and Kehoe *et al.*, 2004, for German and Spanish). In contrast with these studies, the current study examines two stop systems, both with primarily lag voicing on stops. English voiceless stops and Korean aspirated and lenis stops fall into the category of long-lag VOT, and English voiced stops and Korean fortis stops fall into the category of short-lag VOT. Thus, the current study examines language learners who need to partition their VOT range in a fine-grained way if they are to accommodate Korean and English stops.

III. METHODS

A. Participants

The data were collected from 40 adult participants. None of the participants reported being diagnosed with a language or reading disorder. All participants passed a pure-tone hearing test in both ears from 500 to 4000 Hz at octave intervals (38 at 20 dB and 2 at 25 dB). The participants were divided into four groups of 10 each: native monolingual speakers of English (NE group), native monolingual speakers of Korean[1] (NK group), early Korean-English bilinguals (Early group) and late Korean-English bilinguals (Late group). The NK and the bilingual groups all consisted of 7 female and 3 male participants, and the NE group consisted of 6 female and 4 male participants. The English monolinguals were students or affiliates of the University of Oregon (mean age =28.8 years). The Korean monolinguals were all students at the American English Institute at the University of Oregon who had lived, at the time of the experiment, in an English-speaking country for less than 5 months (mean age =24.8 years).

The bilingual participants all spoke Korean as their first language and had lived in the United States for at least 6 years. The Early group had begun learning English between 1 and 6 years, typically when they started school or preschool ($M=3.8$ years). The age of acquisition for the Late group ranged from 15 to 34 years old and was determined by age of first massive exposure to English when they moved to the United States ($M=21.4$ years). Even though the Late group had some exposure to English in middle school and high school in Korea, they had received very limited exposure to native English speakers. Speakers in both bilingual-groups used English regularly on a daily basis at the time of the study.

Although the two groups of bilinguals differed from each other principally in the age at which they were first immersed in English, the two groups differed in other ways as well. The Late group was older at the time of testing (31 vs 20 years) and used English less overall on a daily basis according to self-report (54% vs 81%). Also, their average length of residence in the U.S. was shorter than the Early group, although both groups had been in the U.S. for a considerable amount of time (9.8 vs 18.2 years).

TABLE II. Korean stops recorded for acoustic measurements.

Aspirated	Lenis	Fortis
/pʰɑtɑ/ "to dig"	/pɑtɑ/ "sea"	/p*ɑtɑ/ "to grind"
/tʰɑtɑ/ "to ride"	/tɑto/ "tea ceremony"	/t*ɑtɑ/ "to pick"
/kʰɑdɨ/ "card"	/kɑtɑ/ "to go"	/k*ɑtɑ/ "to peel"

As the goal of the present study was to compare early and late bilinguals, and early bilinguals were likely to be highly proficient in English, it was critical to recruit late bilinguals who were also highly proficient in English. To this end, bilinguals who had long lengths of residence (at least 6 years in the U.S.) and high levels of education (all had at least some college education in the U.S.) were recruited. In addition to this, a standardized test of English proficiency, the Test of Adolescent and Adult Language (TOAL), was administered to the bilinguals. This test was chosen because it was designed to test highly proficient English speakers at adult levels (Hammill *et al.*, 1994). Two of the TOAL subtests, those focusing on listening vocabulary and listening grammar, were administered and scores were recorded as number correct from a low of zero to a high of 35. Performance on the tests was comparable between the two bilingual groups. There was no significant difference between the two bilingual groups for the TOAL Test 1 (25.0 vs 21.1) $[F(1,18)=2.3, p>0.05]$ or Test 2 (27.2 vs 25.3) $[F(1,18)=0.51, p>0.05]$.

B. Speech material

Due to the cross-linguistic nature of this study, the target sounds, English voiceless and voiced stops, and Korean voiceless aspirated, lenis, and fortis stops, were matched for phonological environment in the words to be analyzed: utterance initial, and preceding a low vowel /ɑ/. Tables II and III present the word lists in a phonemic transcription.

Each target word was presented on a flashcard in the orthography of each language in two ways. First, to help facilitate the understanding of the meaning of the words and thus, to record production of known words, a sentence contextualizing the target word was presented. (For example, the contextualizing sentence for "pot" was "We made soup in the pot.") Second, in order to record the target words in a constant prosodic environment, each word was presented in a carrier sentence in an utterance initial position. The English carrier sentence was "__*is the word*," and the Korean one was "[__hɑsɛjo]." ("Say__"). The words in the carrier sentences were analyzed.

Each of the participants was recorded using a high-quality, head-mounted microphone and a Sony DAT recorder in a sound-attenuated room. The speakers read the cards three times in randomized blocks, three for each language (stop types were mixed within each block). The two groups of Korean-English bilinguals (Early and Late) first read the English cards three times and then the Korean cards three times. As control groups, the NE group and the NK group read the English and the Korean cards, respectively. A total of 1350 word productions was analyzed (9 Korean words $\times 3$ repetitions $\times 10$ participants $\times 3$ groups and 6 English words $\times 3$ repetitions $\times 10$ participants $\times 3$ groups).

TABLE III. English stops recorded for acoustic measurements.

Voiceless	Voiced
/pʰɑt/ "pot"	/bɑt/ "bot"
/tʰɑt/ "tot"	/dɑt/ "dot"
/kʰɑt/ "cot"	/gɑt/ "got"

The participants were instructed, in English, to read the cards at a comfortable speaking rate and loudness level and to repeat any utterance when they were unsatisfied with their production. Before the recording session began, speakers were given a chance to rehearse the cards. All speakers reported familiarity with the target words. The utterances were digitally recorded at a sampling rate of 32 000 Hz and digitally transferred to a personal computer as wave files with the same sampling rate using the Kay Elemetrics's COMPUTERIZED SPEECH LAB (CSL). The acoustic measurements were performed using PCQUIRER by SCICON, a speech analysis program.

C. Measurements

1. Voice-onset time (VOT)

The VOT of the initial stop in each target word was measured to the nearest 1 ms from the beginning of the stop-burst release to the onset of the periodic portion of the waveform. The onset of the vowel in the waveform was determined by the onset of the first full glottal pulse of the vowel. The onset of the voicing energy in the second formant shown in a time-locked spectrogram was used to help determine voicing onset in conjunction with the waveform. In the few productions with prevoicing, the VOT was measured as a negative number as the time interval between the onset of periodic pulsing during the closure up to the stop release (Lisker and Abramson, 1964). As is common practice in VOT studies (Lisker and Abramson, 1964; MacKay *et al.*, 2001), the negative values for occasional prevoiced stops were not included in calculating the mean VOT values for a given stop type. The English voiced stops were prevoiced by the NE group 17% of the time and the Early and the Late groups 3% and 0% of the time, respectively. Other stop types were not prevoiced by any of the groups.

2. H1-H2 difference

The amplitudes (dB) for the first (H1) and the second (H2) harmonics were measured at the beginning of the vowel. A Hamming window was centered after the first full glottal pulse in the waveform and a narrow-band FFT spectrum (1024 points) was calculated. Zero padding was used to smooth the spectral peaks. The difference in intensity between H1 and H2 is thought to be an indicator of voice quality (Cho *et al.*, 2002). Breathy voicing is characterized by relatively more energy in the fundamental frequency (measured as H1), with a steep falling off in spectral slope. On the other hand, creaky voicing is produced with relatively more energy in H2 and the higher harmonics. Thus, a greater H1-H2 difference would indicate a more breathy voicing

quality, and a smaller or negative H1-H2 difference would indicate a more creaky quality of the voicing. The values submitted to the statistical analyses were obtained by subtracting the values of H2 from values of H1.

The amplitudes of H1 and H2 are affected by formant frequencies of the vowel (Fant, 1960). Only low vowels were measured; thus, formant frequencies should be relatively comparable across the data set. However, as place of articulation of the preceding stop varied, the formant frequencies at the vowel onset, especially *F*2, would be affected differently by the transition from the consonant into the stable portion of the vowel (Delattre, Liberman, and Cooper, 1955). Thus, some effect of formant transitions on the H1 and H2 measures would be expected for the data. However, this effect does not pose a serious problem to the data analysis presented here: All comparisons had equal numbers of tokens for the three places of articulation in each of the groups being compared. In other words, the effect of formant transitions likely introduced some variance in the data, but the variance was introduced equally for each group being compared.

Another consideration about the H1 and H2 measures relates to gender differences. Studies with English speakers have found that female speakers have a greater difference between H1 and H2 than male speakers, indicating that female speakers tend to have more breathy voicing (Hanson and Chuang, 1999; Klatt and Klatt, 1990). Thus, the female speakers in the current study may have greater H1-H2 values than the male speakers. However, as roughly equal numbers of male and female speakers were included for each of the four participant groups (4/6 for one and 3/7 for the other three), gender differences should not confound comparisons between the groups.

3. Fundamental frequency (F0)

The effect of consonant type on the fundamental frequency (*F*0) of the following vowel was also measured. Due to the common occurrence of irregular glottal pulses from creaky voicing after Korean fortis stops, reliable measures for *F*0 at vowel onset were not possible. Therefore, the effect of a consonant on the following vowel's *F*0 was estimated by taking a measure at vowel midpoint. By vowel midpoint, the creaky voicing found for Korean fortis stops had subsided and it was possible to obtain reliable *F*0 measures.

Fundamental frequency was measured at the temporal midpoint of the vowel by obtaining the frequency of the first harmonic from a narrow-band FFT spectrum (1024 points). The frequency of the glottal pulse at the temporal midpoint of the vowel was also obtained by measuring the duration of the relevant period in milliseconds and dividing it into 1000. This measure was checked against the frequency measured for the first harmonic. When the measures differed, the hand-measured frequency was recorded.

Because *F*0 varies according to each individual, especially across age and gender groups, the measured raw *F*0 values for each participant were normalized. Each *F*0 value produced by a given participant was divided by the mean value of all the midpoint measurements for that participant. Mean values were obtained separately for English stops (voiceless, voiced) and for Korean stops (aspirated, lenis, fortis). Thus, a normalized value more than 1 represents a *F*0 higher than the mean *F*0 for that speaker for that language, and a value less than 1 represents a lower *F*0 than the mean *F*0 for that speaker for that language. These normalized *F*0 values were used for between-group analyses. However, the raw *F*0 values were used for within-group analyses, in which stops produced by the same individuals were compared.

IV. RESULTS

In order to examine age of learning effects on segmental production of English and Korean stops and to examine the interaction of the English and Korean stop systems, two kinds of analyses were conducted. First, between-group analyses were carried out for English and Korean stops separately to investigate differences between the monolingual groups and the early and late bilingual groups in the production of English and Korean stops. Two separate analyses were performed, one for English stops and one for Korean stops. Second, within-group analyses were performed to examine how the stop systems of the two languages were organized in the bilinguals. In these analyses, the difference between the early and late bilinguals in the extent to which the stop systems were held independent from each other was investigated. The alpha level is set at 0.05 for the statistical tests reported here unless otherwise indicated, as in the case of multiple comparisons.

Table IV presents the mean values of the Korean and English stops by group for the acoustic measures analyzed here. The measures of VOT, H1-H2, and normalized *F*0 will be examined in between-group comparisons and the measures of VOT, H1-H2, and raw *F*0 will be examined in within-group comparisons.

A. Age of learning effects on the production of English stops: Between-groups analysis

The results of a three-way, group (NE, Early, Late) by stop type (English voiceless, voiced) by place (labial, alveolar, velar) multivariate repeated measures analysis with the dependent measures of VOT, H1-H2, and normalized *F*0 returned significant main effects for group $[F(6,294)=5.892, p<0.001]$, stop type $[F(3,147)=353.683, p<0.001]$, and place $[F(6,294)=7.455, p<0.001]$. The analysis also revealed interactions of group × stop type $[F(6,294)=7.165, p<0.001]$ and stop type × place $[F(6,294)=2.352, p=0.031]$. The results indicate that the group effect depended on stop type.

MANOVAs on each stop type returned a significant effect of group for English voiceless stops $[F(6,158)=6.225, p<0.001]$ and for English voiced stops $[F(6,132)=6.804, p<0.001]$ (alpha level was adjusted for two tests, $\alpha=0.025$). Pairwise comparisons conducted separately for the English voiceless and voiced stops revealed differences between the Late and the NE groups. In the case of both the voiceless and voiced stops, the Late group was different from the NE group for all three measures. As can be seen in Fig. 1, the Late group produced the voiceless stops differently from the NE group. As shown in Fig. 1(a), the Late group had

TABLE IV. Mean and standard errors in parentheses for the production of Korean and English stops for four acoustic measures: voice-onset time (VOT), the difference between the first two harmonics at the following vowel onset (H1-H2), raw fundamental frequency of the following vowel midpoint (*F*0raw), and normalized fundamental frequency of the following vowel midpoint (*F*0norm; see Sec. III C 3 for method of calculation). The Korean stops were produced by three groups of ten speakers each: early Korean-English bilinguals (Early), late Korean-English bilinguals (Late), and native monolingual speakers of Korean (NK). The English stops were produced by three groups of ten speakers each: Early, Late, and native monolingual speakers of English (NE). The numbers in each group column header indicate the number of male (first) and female (second) speakers.

Stop type	Measure	Early 3/7	Late 3/7	NK 3/7	NE 4/6
Korean	VOT (ms)	82 (3)	79 (4)	68 (3)	
aspirated	H1-H2 (dB)	2.9 (0.5)	2.7 (0.5)	2.3 (0.5)	
	*F*0raw (Hz)	219 (13)	214 (11)	234 (12)	
	*F*0norm	1.113 (0.011)	1.128 (0.017)	1.119 (0.011)	
Korean	VOT (ms)	67 (3)	61 (4)	63 (4)	
lenis	H1-H2 (dB)	1.7 (0.1)	2.3 (0.8)	1.2 (0.6)	
	*F*0raw (Hz)	166 (8)	160 (8)	173 (8)	
	*F*0norm	0.844 (0.010)	0.851 (0.011)	0.844 (0.010)	
Korean	VOT (ms)	14 (1)	16 (2)	11 (1)	
fortis	H1-H2 (dB)	−1.9 (1.4)	−1.2 (0.9)	−3.6 (1.3)	
	*F*0raw (Hz)	205 (13)	194 (11)	216 (11)	
	*F*0norm	1.042 (0.013)	1.021 (0.016)	1.036 (0.008)	
English	VOT (ms)	77 (4)	86 (4)		72 (3)
voiceless	H1-H2 (dB)	−0.2 (0.6)	2.8 (0.6)		0.2 (0.6)
	*F*0raw (Hz)	170 (10)	193 (11)		167 (9)
	*F*0norm	1.036 (0.011)	1.081 (0.017)		1.021 (0.011)
English	VOT (ms)	17 (2)	19 (2)		14 (1)
voiced	H1-H2 (dB)	−1.6 (0.6)	−0.2 (0.7)		−2.9 (0.4)
	*F*0raw (Hz)	160 (9)	162 (9)		161 (8)
	*F*0norm	0.963 (0.012)	0.918 (0.018)		0.979 (0.008)

longer VOT values, as shown in Fig. 1(b), the Late group had greater H1-H2 values, and as shown in Fig. 1(c), the Late group had higher normalized *F*0 values. The Late group also differed from the NE group in the production of the voiced stops. As shown in Fig. 1(a), the Late group had longer VOT values, as shown in Fig. 1(b), the Late group had greater H1-H2 values, and as shown in Fig. 1(c), the Late group had lower normalized *F*0 values.

The Early group, on the other hand, was not significantly different from the NE group for either English voiceless or voiced stops (Tukey's HSD, $p<0.05$). As can be seen in Fig. 1, the Early group's production of the English stops was similar to the NE group's for the three measures examined.

To summarize, the Late group was different from the NE group in the production of English stops for VOT, H1-H2, and *F*0. For the voiceless stops, the VOT was longer, the following vowel had a greater positive difference between the first two harmonics (more breathy phonation), and the normalized *F*0 was higher. For the voiced stops, the VOT was longer, the following vowel had a smaller negative difference between the first two harmonics (less creaky), and the normalized *F*0 was lower. In contrast, the Early group was not different from the NE group for either English voiceless or voiced stops.

B. Age of learning effects on the production of Korean stops: Between-groups analysis

The results of a three-way, group (NK, Early, Late) by stop type (Korean aspirated, lenis, fortis) by place (labial, alveolar, velar) multivariate repeated measures analysis with dependent measures of VOT, H1-H2, and normalized *F*0 returned significant main effects for group [$F(6,482)=3.182, p=0.005$], stop type [$F(6,482)=383.125, p<0.001$], and place [$F(6,482)=9.761, p<0.001$]. There was no interaction between these factors ($p>0.05$). These results indicate that the groups differed from one another in their production of Korean stops in a consistent manner across stop type. To address the research question of whether there were differences in the production of Korean stops by age of English acquisition, planned comparisons investigating the effect of group for each stop type were conducted.

MANOVAs on each stop type returned a significant effect of group for Korean fortis stops only [$F(6,158)=4.607, p<0.001$]. The effect of group was not significant for the aspirated and lenis stops (alpha level was adjusted for three tests, $\alpha=0.016$). Pairwise comparisons (Tukey's HSD tests, $p<0.05$) investigating the three measures separately for the Korean fortis stops revealed a difference between the groups for VOT only. The Late group produced the fortis stops with longer VOT values than the NK group. As can be seen in Fig. 2(a) on the right-hand side, the VOT produced by the Late group for the fortis stops was only slightly longer, but the difference was significant. Although there was not a significant effect of group for the MANOVA investigating the Korean aspirated stops, as can be seen in Fig. 2(a) left-hand side, both the Early and Late groups tended to have longer VOT values than the NK group. On the other hand,

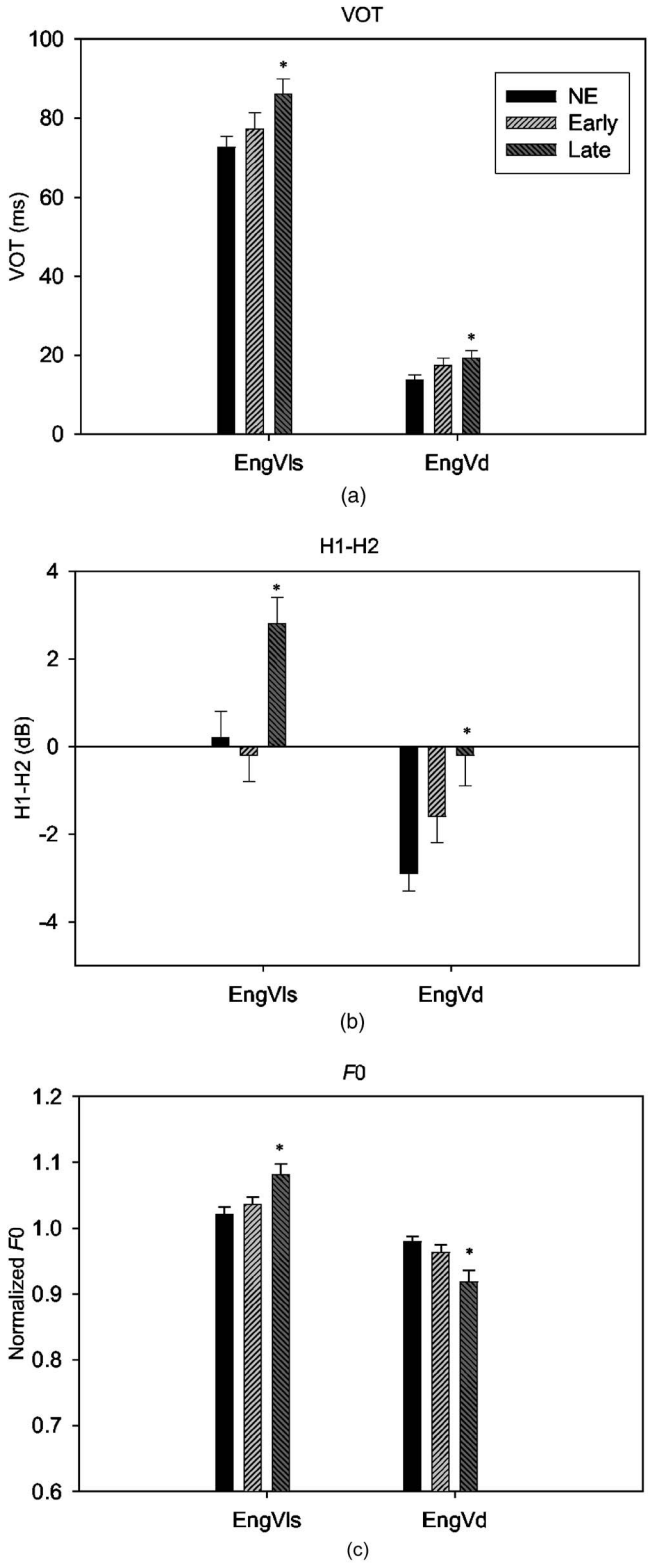


FIG. 1. Mean values with standard errors for the production of English stops (voiceless and voiced) for three acoustic parameters [(a) VOT; (b) H1-H2; (c) normalized *F*0] by three groups of ten speakers each (NE, early bilinguals, and late bilinguals). *=statistically different from the NE group.

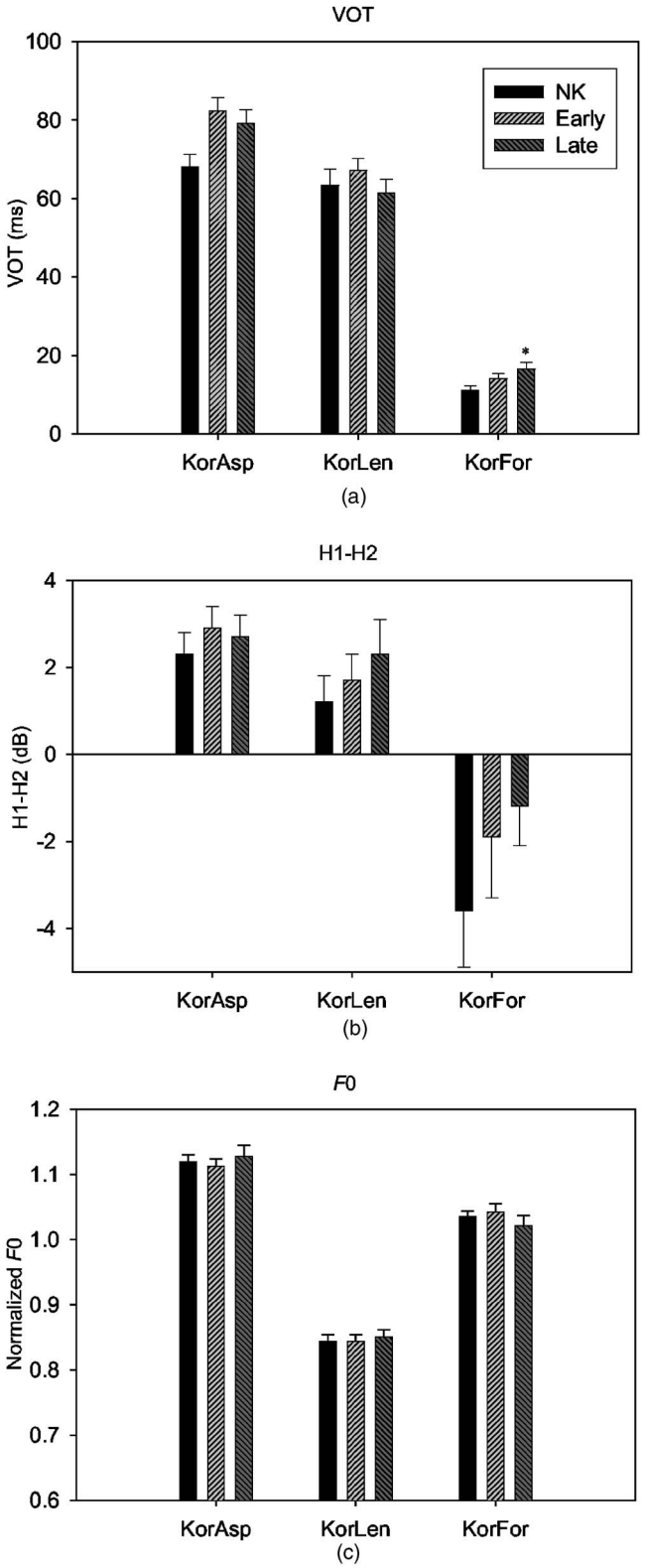


FIG. 2. Mean values with standard errors for the production of Korean stops (aspirated, lenis, and fortis) for three acoustic parameters [(a) VOT; (b) H1-H2; (c) normalized *F*0] by three groups of ten speakers each (NK, early bilinguals, and late bilinguals). *=statistically different from the NK group.

there was no significant difference between the three groups for the H1-H2 and normalized *F*0 measures for any of the three stop types. As can be seen in Fig. 2(b), there were no clear group differences for the H1-H2 measure. As can be seen in Fig. 2(c), there was very little difference between the groups for the normalized *F*0 values.

C. Age of learning effects on the degree of independence between English and Korean stop systems: Within-group analyses

This section addresses the question of whether age of learning differences between early and late bilinguals affect

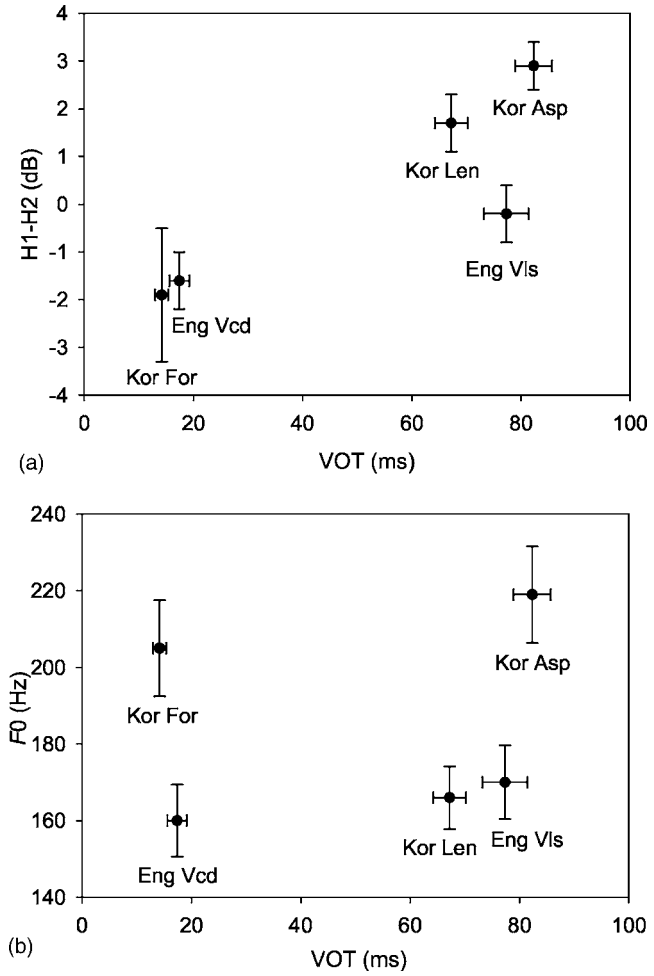


FIG. 3. The early bilingual group's production of English (voiceless, voiced) and Korean (aspirated, lenis and fortis) stops. Mean values with standard errors for three acoustic parameters (VOT, H1-H2, raw *F*0). The values for H1-H2 and VOT are shown in (a) and *F*0 and VOT are shown in (b).

the extent to which their two phonological systems are held independent from each other. To this end, the grouping patterns of the five stop categories of Korean and English were examined as a function of age of learning. For each group, ten pairwise comparisons were made: one comparison for the English stop types, three comparisons for Korean stop types, and six combined comparisons for English and Korean stop types. The results of this analysis indicate the number of independent stop types for each group.

1. Early group

Ten multivariate pairwise comparisons (i.e., all possible combinations of the English and Korean stops) with the dependent variables of VOT, H1-H2, and raw *F*0 were made. For the Early group, all ten comparisons were significantly different ($p \leqslant 0.005$, alpha level was adjusted to 0.005 for ten tests). These results suggest that the English and the Korean stop categories were separated from one another within their own stop systems and, at the same time, the five stop categories of Korean and English were all separated from one another.

Figure 3 displays the VOT, H1-H2, and raw *F*0 values for the five stops produced by the Early group (see Table IV for exact values). The three Korean stops are clearly distinguished from each other by the three measures. In Fig. 3(a), note that the three Korean stops are well-distinguished from each other in terms of VOT (horizontal axis) and that the Korean fortis stops are set apart in terms of H1-H2 (vertical axis). In Fig. 3(b), note that the Korean lenis stops have a much lower *F*0 (vertical axis) than the other two Korean stops. Moving to the English stops, as shown in Fig. 3(a), the two English stops are distinguished from each other primarily by VOT, and secondarily by H1-H2. As shown in Fig. 3(b), the English stops are not well-differentiated by *F*0. As for the Korean-English stop comparisons, the Korean fortis stops, as shown in Fig. 3(b), are primarily distinguished from the English voiced stops by *F*0. The Korean aspirated stops, as shown in Figs. 3(a) and 3(b), are distinguished from English voiceless stops by H1-H2 and *F*0. The Korean lenis stops, as shown in Fig. 3(a), are distinguished from the English voiceless stops primarily by H1-H2 and secondarily by VOT.

In summary, the Early bilingual group had five distinct stop types in their combined Korean English systems. English voiceless, English voiced, Korean aspirated, Korean lenis, and Korean fortis stops each formed a group that could be statistically distinguished from the other stop types.

2. Late group

The multivariate pairwise comparisons with the dependent variables of VOT, H1-H2, raw *F*0 for the Late group revealed no significant differences between the English voiced and Korean fortis stops and between the English voiceless and Korean aspirated stops. The other eight comparisons returned significant differences for the pairing of the other stop types ($p < 0.005$). These results suggest that, in the stop system of the late bilinguals, the English stop categories are distinguished from each other and the Korean stops are also distinguished from each other. However, the English voiced and Korean fortis stops, on the one hand, and the English voiceless and Korean aspirated stops, on the other hand, are not separated from each other.

Figure 4 displays the VOT, H1-H2, and raw *F*0 values for the five stops produced by the Late group (see Table IV for exact values). The three Korean stops are clearly distinguished from each other by the three measures. As shown in Fig. 4(a), the Korean fortis stops have lower H1-H2 values than the other Korean stops and all three Korean stops are distinguished by VOT. As shown in Fig. 4(b), the Korean lenis stops have a lower *F*0 than the other Korean stops. The two English stops are also clearly distinguished from each other by the three measures: The English voiced stops have lower H1-H2 and VOT values [see Fig. 4(a)], as well as lower *F*0 values [see Fig. 4(b)].

In contrast, the Korean aspirated stops are not well-distinguished from the English voiceless stops by any of the three measures. Note the very close H1-H2 and VOT mean values in Fig. 4(a) and overlapping distributions of *F*0 values in Fig. 4(b). The Korean fortis and English voiced stops, though not significantly different from each other in the MANOVA reported above, do exhibit some separation with

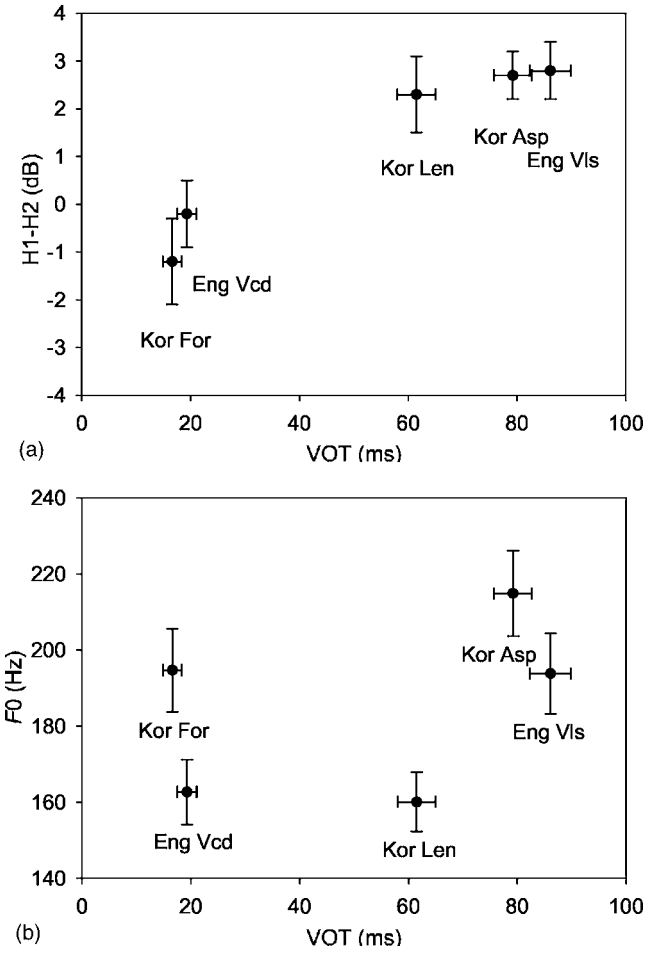


FIG. 4. The late bilingual group's production of English (voiceless, voiced) and Korean (aspirated, lenis and fortis) stops. Mean values with standard errors for three acoustic parameters (VOT, H1-H2, raw *F*0). The values for H1-H2 and VOT are shown in (a) and *F*0 and VOT are shown in (b).

regard to *F*0 [see Fig. 4(b)]. Specifically, the English voiced stops have lower *F*0 values, which are similar to the Korean lenis stops. Thus, the English voiced stops are similar to the Korean fortis stops in terms of VOT and H1-H2 [see Fig. 4(a)] but similar to the Korean lenis stops in terms of *F*0.

In summary, the Late bilingual group had three distinct stop types in their combined Korean-English systems. English voiceless and Korean aspirated stops formed one group. English voiced and Korean fortis stops formed another group, though demonstrating a trend towards different *F*0 values. Finally, the Korean lenis stops formed a group by themselves.

V. DISCUSSION

The results reported here suggest that there exist age of learning effects between the early bilinguals and the late bilinguals in the production of English and Korean stops and in the extent of independence between the Korean and English stop systems. The early bilinguals were not different from the English monolinguals for either English voiced or voiceless stops, nor were they different from the Korean monolinguals for any of the three Korean stops. At the same time, the early bilinguals maintained five distinct stop types in their stop systems (see Fig. 3). Thus, the early bilinguals seem to have established English voiced and voiceless stop categories independent from the Korean stop categories. In other words, there was no apparent interaction between the L1 and L2 stop system, as all productions were monolingual-like.

In contrast, the late bilinguals produced English voiceless and voiced stops differently from English monolinguals in terms of VOT, as well as H1-H2 and *F*0 of the following vowel. They also produced Korean fortis stops slightly differently from Korean monolinguals in terms of VOT. Whether or not these differences are perceptually salient to native monolingual speakers of the respective languages is an open question. Further research will be needed to make this determination. Nonetheless, the productions of the late bilinguals were found to be significantly different from monolingual speakers in this investigation.

Thus, the late bilinguals were found to produce both English and, to a lesser extent, Korean stops differently than monolingual speakers. At the same time, the late bilinguals had fewer distinct stop types in their combined Korean-English stop system. Their Korean aspirated and English voiceless stops were not distinguished from each other in production by any of the three acoustic measures (see Fig. 4). Their English voiced stops did not form a distinct category, but exhibited similarities with both the Korean fortis and lenis stops (see Fig. 4).

The apparent merger between Korean aspirated and English voiceless stops did not seem to influence the production of Korean aspirated stops, but did seem to influence the production of English stops. As can be seen in Fig. 1 and Table IV, the English voiceless stops produced by the late bilinguals had longer VOT values (86 ms) than the English monolinguals (72 ms), a greater difference between the first two harmonics of the following vowel (2.8 dB) than the English monolinguals (0.2 dB), and higher normalized *F*0 values of the following vowel (1.08) than the native English monolinguals (1.02). The values for the late bilinguals' English voiceless stops were similar to the values for their Korean aspirated stops (VOT 79 ms, H1-H2 2.7 dB, and normalized *F*0 values 1.12; see Fig. 2). Additionally, the late Korean bilinguals did not produce Korean aspirated stops differently from the Korean monolinguals. Thus, it seems that the late bilinguals were using their Korean aspirated stops when speaking English. The L1 category seemed to be used for both the L1 and L2, with no effect on L1 pronunciation.

The status of English voiced stops produced by the late bilinguals is not as clear as the status of voiceless stops. While it seems likely that the late bilinguals had transferred the production characteristics of Korean aspirated stops to English voiceless stops, this same group's production of English voiced stops is not fully equivalent to either Korean fortis or Korean lenis stops (see Fig. 4). On the one hand, the values for VOT and H1-H2 of the following vowel are similar for English voiced and Korean fortis stops (19 vs 16 ms and −0.2 vs −1.2 dB, respectively). On the other hand, the values for *F*0 of the following vowel are similar for English voiced and Korean lenis stops (162 vs 160 Hz, respectively).

The incorporation of the English voiced stop into the phonological system of the late bilinguals may have influ-

enced the production of the Korean fortis stops. The late bilinguals produced Korean fortis stops differently from the Korean monolinguals. The longer VOT found for the Korean fortis stops produced by the late bilinguals (16 ms) is more like the VOT of English voiced stops produced by the English monolinguals (14 ms) than that produced for the Korean fortis stops by the Korean monolinguals (11 ms).[2]

The finding that the late bilinguals had merged L1 and L2 stops to varying degrees can be interpreted in light of models of L2 speech learning. Theoretical models have addressed the role of perceptual similarities between L1 and L2 sounds in second language speech learning. The Speech Learning Model (SLM) (Flege, 1995) posits that the greater the perceived phonetic distance between an L2 and L1 segment, the more likely that differences between these segments will be discerned and, thus, the more likely that an independent phonetic category for the L2 sound will eventually be established. Alternatively, if instances of an L2 category continue to be identified as instances of an L1 category, a "merged" category may develop over time. Similarly, the Perceptual Assimilation Model (PAM) (Best, 1995a, 1995b) predicts that some L2 segments will be more easily distinguished than others and that differences in similarity may ultimately affect L2 speech learning.

In a cross-language mapping experiment, Schmidt (1996) investigated Korean speakers' identification of English consonants in terms of Korean consonant categories. The study revealed that the native Korean speakers consistently labeled English voiceless stops as Korean aspirated stops with high similarity ratings. It also revealed that English voiced stops were labeled with roughly equal frequency as both Korean lenis stops and fortis stops and had lower similarity ratings than those between Korean aspirated and English voiceless stops. Thus, considering the findings of the present study, these cross-language mapping patterns may provide an explanation for the differences in acquisition of English voiceless and voiced stops for the late bilinguals. As mentioned earlier, the late bilinguals apparently used Korean aspirated stops to produce English voiceless stops. In the case of English voiced stops, their production was similar to their Korean fortis stops in terms of VOT and H1-H2 of the following vowel but similar to Korean lenis stops in terms of *F*0 of the following vowel. The apparent merger of the Korean aspirated stops and English voiceless stops may be because the late bilinguals identified instances of English voiceless stops as Korean aspirated stops and, therefore, established a fully merged L1-L2 category with the production characteristics of Korean aspirated stops. The less-complete merger of English voiced stops with any single Korean stop type may be due to a less-clear cross-language mapping between English voiced stops and a single Korean stop type. In other words, since the L2-L1 mapping for English voiced and Korean stops was of a two-category type (Best, 1995b), in that English voiced stops were mapped to both Korean fortis and Korean lenis stops, a fully merged category was not developed. Instead, English voiced stops may have developed as something of a hybrid between the two Korean stop types.

Let us now turn to the finding that the early bilinguals maintained independence between their two stop systems, whereas the late bilinguals evidenced a merger between the systems. Previous research has suggested that the degree of separation between the languages may be affected by age of L2 acquisition. Increased delays in L2 exposure have been related to increased influence between the L1 and L2 phonological systems. It has been well demonstrated that later learners show more L1 effects on L2 in segmental speech production than early learners (e.g., Flege, 1991; MacKay *et al.*, 2001; Piske *et al.*, 2002). Additionally, emerging evidence suggests that earlier bilinguals may have greater independence between their two phonological systems than late bilinguals. Later bilinguals may be more likely to merge L1 and L2 phonetic categories, whereas earlier bilinguals may be more likely to acquire new categories for the L2 (Flege, 1987a, 1987b, 1995).

However, the extent to which the two phonological systems of early bilinguals are separated is still not clear. Three previous studies have found non-monolingual-like production of either the L1 or L2 that was attributed to dissimilatory effects in a combined L1-L2 phonological system. Guion (2003) found that early bilinguals demonstrated an effect of the combined L1 and L2 systems on the production of L1 vowels in Quichua-Spanish bilinguals. Flege *et al.* (2003) also found an influence of the combined L1 and L2 systems on the production of L2 vowels in early Italian-English bilinguals. Finally, Harada (2003) found an effect on L1 stop production in early Japanese-English bilinguals that could be attributed to a mutual interaction of the L1 and L2 systems. On the other hand, in the current study, the early Korean-English bilinguals' stop production evidenced no influence from either language. In other words, they produced both the L1 and L2 stops like the monolingual speakers. There was no evidence of any sort of dissimilation between the stops in a combined L1-L2 system.

Given the results of the Guion (2003) study that early but not simultaneous bilinguals showed effects of merged phonological systems, the prediction was made that the younger the L2 learner, the more likely it would be for independent phonological systems to be developed. The early learners in the current study had an earlier age of acquisition than the early learners in the three studies mentioned above. In the current study, the average age of acquisition was 3.8 years (range 1–6). The early bilinguals reported by Guion had an average age of acquisition of 5.8 years (range 5–7). The early bilinguals reported by Flege *et al.* (2003) had an average age of acquisition of 7 years (range 2–13). The early bilinguals reported by Harada (2003) had an average age of acquisition of 5 years (range 4–6) (Harada, 2005 p.c.). Thus, the discrepancy between the current study and those reported previously may be due to variation in age of acquisition. Perhaps the onset of second language acquisition by the bilinguals in this study enabled the development of more fully independent phonological systems than was the case for the later onset of second language acquisition in the other studies.

A likely explanation for age effects on the separation of the L1 and L2 phonological systems may be found in an

"emergentist" view developed for general second language acquisition. As summarized in Hernandez, Li, and MacWhinney (2005), an "emergentist" model predicts that the degree of separation between a bilingual's languages is related to the development, or entrenchment, of the first language and how that entrenchment affects the acquisition of a second language. An entrenched language exhibits the property of resonance. In other words, a given language form can activate other language forms within the same language in a resonant interaction. When a new language is being learned, it is dependent on the L1 in a parasitic way in that it resonates with the L1. If the L2 is to be fully independent, it will have to set up its own language-internal resonance and not interact with the L1. It is proposed that the ability to set up a new resonant system is affected by the level of entrenchment of the first language. Thus, it is predicted that earlier, balanced bilinguals should have greater levels of separation between their two languages than later bilinguals.

The findings of the present study are consistent with these predictions. The early bilinguals established L2 phonetic categories for English stops independently from Korean stops, whereas the late bilinguals had merged categories of Korean and English stops. Thus, the findings of the present study suggest that age-related differences in the establishment, or entrenchment, of the L1 may affect the reorganization process of bilinguals' phonological systems.

ACKNOWLEDGMENTS

This research was supported by a grant (DC 05132) from the National Institutes of Health (National Institute on Deafness and Other Communication Disorders). The authors thank J. J. Clark for her help with data collection and Melissa Redford, Jonathan Loftin, two anonymous reviewers, and Brad Story for their helpful comments.

[1]The monolingual Korean had studied English in middle school and high school in Korea. However, they never used English on daily basis and did not have a functional command of English.

[2]The bilingual groups also tended to have longer VOT for Korean aspirated stops than the monolingual group. It is unclear, however, if this tendency is related to the acquisition of English. Silva (2002; 2004) has reported that young Korean speakers, such as those recorded for our Korean monolingual group, have shortened the VOT produced for aspirated stops. Further investigation with older monolingual speakers of Korean, who maintain longer VOT for aspirated stops, is warranted before drawing any conclusions.

Abramson, A. S., and Lisker, L. (**1985**). "Relative power of cues: *F*0 shift versus voice timing," in *Phonetic linguistics: Essays in Honor of Peter Ladefoged*, edited by V. Fromkin (Academic, Orlando), pp. 25–33.

Ahn, H. (**1999**). "Post-release phonatory processes in English and Korean: Acoustic correlates and implications for Korean phonology," Doctoral dissertation, University of Texas at Austin.

Best, C. T. (**1995a**). "Learning to perceive the sound pattern of English," in *Advances in Infancy Research*, edited by L. P. Lipsitt and C. Rovee-Collier (Ablex, Norwood, NJ), Vol. **9**, pp. 217–304.

Best, C. T. (**1995b**). "A direct realist perspective on cross-language speech perception," in *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, edited by W. Strange (York, Timonium, MD), pp. 171–204.

Cho, T., Jun, S.-A., and Ladefoged, P. (**2002**). "Acoustic and aerodynamic correlates of Korean stops and fricatives," J. Phonetics **30**, 193–228.

Delattre, P. C., Liberman, A. M., and Cooper, F. S. (**1955**). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. **27**, 769–773.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Flege, J. E. (**1987a**). "Effects of equivalent classification on the production of foreign language speech," in *Sound Patterns in Second Language Acquisition*, edited by A. James and J. Leather (Foris, Dordrecht), pp. 9–39.

Flege, J. E. (**1987b**). "The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification," J. Phonetics **15**, 47–65.

Flege, J. E. (**1991**). "Age of learning affects the authenticity of voice-onset time in stop consonant production in a second language," J. Acoust. Soc. Am. **89**, 395–411.

Flege, J. E. (**1995**). "Second-language speech learning: theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, edited by W. Strange (York Press, Timonium, MD), pp. 233–277.

Flege, J. E., and Eefting, W. (**1986**). "Linguistic and developmental effects on the production and perception of stop consonants," Phonetica **43**, 155–171.

Flege, J. E., and Eefting, W. (**1987a**). "Production and perception of English stops by native Spanish speakers," J. Phonetics **15**, 67–83.

Flege, J. E., and Eefting, W. (**1987b**). "Cross-language switching in stop consonant perception and production by native Dutch speakers of English," Speech Commun. **6**, 185–202.

Flege, J. E., and Eefting, W. (**1988**). "Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation," J. Acoust. Soc. Am. **83**, 729–740.

Flege, J. E., and Hillenbrand, J. (**1984**). "Limits on phonetic accuracy in foreign language speech production," J. Acoust. Soc. Am. **76**, 708–721.

Flege, J. E., Mackay, I. R. A., and Meador, D. (**1999**). "Native Italian speakers' perception and production of English vowels," J. Acoust. Soc. Am. **106**, 2973–2987.

Flege, J. E., Munro, M. J., and Mackay, I. R. A. (**1995**). "Effects of age of second language learning on the production of English consonants," Speech Commun. **16**, 1–26.

Flege, J. E., Schirru, C., and MacKay, I. R. A. (**2003**). "Interaction between the native and second language phonetic systems," Speech Commun. **40**, 467–491.

Guion, S. G. (**2003**). "The vowel system of Quichua-Spanish bilinguals: Age of acquisition effects on the mutual influence of the first and second language," Phonetica **60**, 98–128.

Hammill, D. D., Brown, V. L., Larsen, S. C., and Weiderholt, J. L. (**1994**). *Test of Adolescent and Adult Language*, 3rd ed. (Pro-ed, Austin, TX).

Han, M. S., and Weitzman, R. S. (**1970**). "Acoustic features of Korean /P,T, K//p,t, k/ and /p^h,t^h,k^h/," Phonetica **22**, 112–128.

Hanson, H. M., and Chuang, E. S. (**1999**). "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," J. Acoust. Soc. Am. **106**, 1064–1077.

Harada, T. (**2003**). "L2 influence on L1 speech in the production of VOT," in Proceedings of 15th International Conference of Phonetic Science, Barcelona, Spain, pp. 1085–1088.

Hernandez, A., Li, P., and MacWhinney, B. (**2005**). "The emergence of competing modules in bilingualism," Trends in Cognitive Science **9**, 220–225.

Kagaya, R. (**1974**). "A fiberscopic and acoustic study of Korean stops, affricates, and fricatives," J. Phonetics **2**, 161–180.

Kehoe, M. M., Lleó, C., and Rakow, M. (**2004**). "Voice onset time in bilingual German-Spanish children," Bilingualism: Lang. Cognit. **7**, 71–88.

Khattab, G. (**2002**). "VOT production in English and Arabic bilingual and monolingual children," in *Perspectives on Arabic Linguistics XIII-XIV*, edited by D. E. Parkinson and E. Benmamoun (John Benjamins, Amsterdam), pp. 1–38.

Khattab, G. (**2003**). "Age, input, and language mode factors in the acquisition of VOT by English-Arabic bilingual children," in Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain, pp. 1–4.

Kim, C.-W. (**1965**). "On the autonomy of the tensity feature in stop classification (with special reference to Korean stops)," Word **21**, 339–359.

Kim, M.-R. Cho. (**1994**). "Acoustic characteristics of Korean stops and perception of English stop consonants," Doctoral dissertation, University of Wisconsin at Madison.

Kim, M.-R., Beddor, P. S., and Horrocks, J. (**2002**). "The contribution of consonantal and vocalic information to the perception of Korean initial stops," J. Phonetics **30**, 77–100.

Klatt, D. H. (**1975**). "Voice onset time, frication and aspiration in word-initial consonant clusters," J. Speech Hear. Res. **18**, 686–705.
Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.
Lehiste, I., and Peterson, G. E. (**1961**). "Some basic considerations in the analysis of intonation," J. Acoust. Soc. Am. **33**, 419–425.
Lisker, L., and Abramson, A. S. (**1964**). "A cross-language study of voicing in initial stops: Acoustical measurement," Word **20**, 384–422.
Mack, M. (1990). "Phonetic transfer in a French-English bilingual child," in *Attitudes and Language Conflict*, edited by P. H. Nelds (Dümmler, Bohn), pp. 107–124.
Mackay, I. R. A., Flege, J. E., Piske, T., and Schirru, C. (**2001**). "Category restructuring during second-language acquisition," J. Acoust. Soc. Am. **110**, 516–528.
Munro, M. J., Flege, J. E., and Mackay, I. R. A. (**1996**). "The effects of age of second-language learning on the production of vowels," Appli. Psycholing. **17**, 313–334.
Ohde, R. N. (**1985**). "Fundamental frequency correlates of stop consonant voicing and vowel quality in the speech of preadolescent children," J. Acoust. Soc. Am. **78**, 1554–1561.
Piske, T., Flege, J. E., Mackay, I. R. A., and Meador, D. (**2002**). "The production of English vowels by fluent early and late Italian-English bilinguals," Phonetica **59**, 49–71.
Schmidt, A. M. (**1996**). "Cross-language identification of consonants. I. Korean perception of English," J. Acoust. Soc. Am. **99**, 3201–3211.
Shimizu, K. (**1996**). *A Cross-language Study of Voicing Contrasts of Stop Consonants in Asian Languages* (Seibido, Tokyo).
Silva, D. (**1992**). "The phonetics and phonology of stop lenition in Korea," Doctoral dissertation, Cornell University, Ithaca, NY.
Silva, D. (**2002**). "Consonant aspiration in Korean: A retrospective," in *Pathways into Korean Language and Culture: Essays in Honor of Young-Key Kim-Renaud*, edited by S.-O. Lee and G. Iversion (Pagijong, Seoul), pp. 447–469.
Silva, D., Choi, Y.-J., and Kim, J. E. (**2004**). "Evidence for a diachronic shift in VOT values for Korean stop consonants," in *Harvard Studies in Korean Linguistics X.S*, edited by S. Kuno, I.-H. Lee, J. Whitman, J. Maling, Y.-S. Kang, and Y.-j. Kim (Harvard University, Department of Linguistics, Cambridge, MA), pp. 173–185.
Sundara, M., Polka, L., and Baum, S., (**in press**). "Production of coronal stops by adult bilingual first language learners of Canadian English and Canadian French: Language-specific and language-general constraints," Bilingualism: Lang. Cognit..
Williams, L. (**1977**). "The perception of stop consonant voicing by Spanish-English bilinguals," Percept. Psychophys. **21**, 289–297.
Williams, L. (**1979**). "The modification of speech perception and production in second-language learning," Percept. Psychophys. **26**, 95–104.

Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese

John S. Pruitt
Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

James J. Jenkins[a)]
Department of Psychology, University of South Florida, 4200 Fowler Avenue, Tampa, FL 33620

Winifred Strange
Ph.D. Program in Speech and Hearing Sciences, Graduate School and University Center, City University of New York, New York, New York 10016

(Received 18 April 2005; revised 12 November 2005; accepted 18 November 2005)

Perception of second language speech sounds is influenced by one's first language. For example, speakers of American English have difficulty perceiving dental versus retroflex stop consonants in Hindi although English has both dental and retroflex allophones of alveolar stops. Japanese, unlike English, has a contrast similar to Hindi, specifically, the Japanese /d/ versus the flapped /r/ which is sometimes produced as a retroflex. This study compared American and Japanese speakers' identification of the Hindi contrast in CV syllable contexts where C varied in voicing and aspiration. The study then evaluated the participants' increase in identifying the distinction after training with a computer-interactive program. Training sessions progressively increased in difficulty by decreasing the extent of vowel truncation in stimuli and by adding new speakers. Although all participants improved significantly, Japanese participants were more accurate than Americans in distinguishing the contrast on pretest, during training, and on posttest. Transfer was observed to three new consonantal contexts, a new vowel context, and a new speaker's productions. Some abstract aspect of the contrast was apparently learned during training. It is suggested that *allophonic* experience with dental and retroflex stops may be detrimental to perception of the new contrast. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2161427]

PACS number(s): 43.70.Kv, 43.71.Hw, 43.71.An [ARB] Pages: 1684–1696

I. INTRODUCTION

Over the last 30 years many studies have shown that adult speakers of any given language have difficulty distinguishing at least some non-native speech sounds (Flege, 1984, 1989; Gillette, 1980; Goto, 1971; Jamieson and Morosan, 1986, 1989; Lisker and Abramson, 1970; Miyawaki *et al.*, 1975; Polka, 1991; Tees and Werker, 1984; Trehub, 1976; for a general review see Strange, 1995). Problems occur in both perception and production of such sounds, and may persist even after years of immersion in the non-native language environment (Best and Strange, 1992; Flege and Eefting, 1987; MacKain *et al.*, 1981; Yamada, 1995; Yamada *et al.*, 1994). Many factors can affect how a speaker of one language perceives the speech sounds of another language and not all non-native contrasts are equally difficult to perceive or produce (Best *et al.*, 1988; Tees and Werker, 1984). Large differences have been found in non-native perception that result from such factors as native language experience, cognitive abilities, motivation of the listeners, the task used to measure perceptual ability, and many variables concerning the stimuli to be perceived (Strange, 1992).

The present study examined one aspect of the difficulty in cross-language perception by comparing the perception of a consonant place contrast in one language (Hindi) by speakers of two other languages (American English and Japanese) in which these consonants do not occur phonemically. What makes this comparison interesting is the different relationship of the phonemic structure of each language to the structure of the target language. This difference is examined in the following.

In the present study, American and Japanese speakers' perception of Hindi dental versus retroflex initial stop consonants was examined using a pretest-training-posttest design with several tests of generalization of learning. For clarity, the research is presented in two parts. First, a test of the perception of the Hindi consonants (without training) is reported. Second, the training program is presented and evaluated both as to the direct effect of the program and its transfer to new materials and speakers. In each part, several questions are addressed. In the pretest, the questions are: (1) Do speakers from the two different language systems differ in their accuracy in identifying the Hindi dental-retroflex contrast? (2) Does the relative difficulty in perceiving this contrast in different consonantal and vowel contexts differ for the two sets of participants? (3) Are the observed differences plausibly related to the phonetic relationships between the languages? For the training study, the questions are: (1) Does training improve perception of this distinction for both American and Japanese speakers? and (2) Does the training transfer to new contexts and new speakers?

[a)]Electronic mail: j3cube@aol.com

 0001-4966/2006/119(3)/1684/13/$22.50

II. PERCEPTION OF THE HINDI DENTAL-RETROFLEX CONTRAST BY NATIVE SPEAKERS OF AMERICAN ENGLISH AND JAPANESE

The long-term goal of this research is to explain the role of native language experience in non-native language speech perception. To accomplish this, one needs to know just which of the many similarities and differences between the two languages are relevant. Jenkins and Yeni-Komshian (1995) discussed a variety of possibilities: differences in systematic phonology, phonetic similarities, allophonic variations, and distinctive features. However, they found no clear answer in the research available at that time.

Most cross-language studies include only two language groups, a target language with some particular phonemic contrast (e.g., /r/ vs /l/ in American English) and a group of listeners whose language does not include such a contrast (e.g., Japanese). Such studies are instructive but leave open the question as to the sources of the problem. By comparing the perceptual performance of two groups of participants with different native languages, neither of which contains the target contrast but which have different phonetic relationships to the target language, it should be possible to begin to tease apart the important dimensions of native language experience that affect non-native phoneme perception.

Polka (1992) examined the perception of the Salish velar /k′/ versus uvular /q′/ glottalized stop contrast by native speakers of English and Farsi. The Farsi language includes a velar-uvular place-of-articulation distinction but it occurs only in voiced plosive stops, not voiceless stops. Farsi also has a voiceless uvular fricative /χ/, but it does not contrast with a voiceless velar fricative /x/. English, on the other hand, does not have any phones produced at the uvular place of articulation. Thus, Polka predicted that Farsi speakers would differentiate the Salish contrast more readily than English speakers. Surprisingly, she found that experience with a presumably related speech contrast, but one that occurred in a different voicing or manner class, did not aid in perceiving that non-native contrast. Examining this in more detail, Polka found a great deal of individual variability among Farsi participants which was correlated with their reports of perceptual assimilation. Farsi participants who indicated that the Salish consonants were similar to Farsi consonants were much more accurate in identification. This supports the theory of Best and her colleagues that differences in perceptual assimilation are predictive of non-native perception (Best, 1993, 1994; Best *et al.* 1988).

More recently, Hardison (2003) studied the perception and production of the American /r/-/l/ distinction by native speakers of Japanese and Korean who were described as "intermediate level learners of American English." Japanese lacks an /r-l/ contrast but Japanese has an "r-like" apical flap occurring in intervocalic and initial positions. Korean has no initial approximate of any sort but has a "voiced alveodental flap intervocalically" and a "voiced apicoalveolar nonvelarized or clear /l/ in syllable-final position" (Hardison, 2003, p 496). One of Hardison's rich set of findings in an extensive pretest-training-posttest study was that Japanese participants had the greatest difficulty with the American contrast in initial singletons and in initial consonant clusters, while Korean participants have their greatest difficulty with the distinction in final singletons and final clusters. Hardison concluded that these difficulties, which persist through training, show some influence from the first language phonology of the participants.

A. Linguistic background

The present study examined a relationship among three language groups, Hindi (the target language), Japanese, and American English. The languages were selected because Hindi contains a contrast (dental versus retroflex place of articulation) which does not appear in the phonemic inventory of English or Japanese (Best and Strange, 1992; Bhat, 1974; Bloch, 1950; Delattre, 1966; Fairbanks and Misra, 1966; Hattori, 1984; Jones, 1964; Jorden, 1963; Martin, 1963; Ogawa, 1982; Polka, 1991; Price, 1981; Shibatani, 1990; Tees and Werker, 1984). However, English and Japanese differ in their phonetic relationship to Hindi with regard to this contrast.

1. Hindi

Hindi makes extensive use of the dental versus retroflex place-of-articulation.[1] It occurs in five pairs of stop consonants: voiceless-unaspirated, voiceless-aspirated, voiced-unaspirated, voiced-aspirated (breathy voiced), and nasal (Fairbanks and Misra, 1966). The present study was limited to syllable-initial consonants because preliminary data concerning post vocalic consonants indicated that they did not pose the same degree of perceptual difficulty for non-native listeners (probably due to extensive coloration of the preceding vowel by retroflex consonants). Because the nasal retroflex does not occur in syllable-initial position, that contrast was not used in this study.

Production studies of retroflexion have shown that the place of articulation can vary greatly (from pre-alveolar to almost post-palatal) and so can the degree of curling of the tongue (Dave, 1977; Dixit, 1990; Dixit and Flege, 1991; Firth, 1957; Ladefoged and Bhaskararao, 1983; Svarny and Zvelebil, 1955). This variation in production occurs across speakers, phonetic contexts, dialects, and languages. In particular, vowel context has been shown to affect the degree of retroflexion. Front vowels lessen retroflexion while back vowels increase it (Dave, 1977; Dixit, 1990; Dixit and Flege, 1991). For this and other reasons, the acoustic consequences of retroflexion are not well understood (Tees and Werker, 1984; Werker and Tees, 1984). Most acoustic studies show differences between dental and retroflex stops as being primarily in the burst and the transition of the second, third, and fourth formants. Retroflex segments contain a lower frequency burst spectrum, which is usually higher in amplitude, and a bunching or clustering of F2, F3, and F4 which rapidly move toward the appropriate vowel (Dave, 1977; Ladefoged and Bhaskararao, 1983; Polka, 1991; Ramasubramanian and Thosar, 1971; Stevens and Blumstein, 1975; see also Jongman *et al.*, 1985). Prevoicing is typical for the Hindi voiced stop consonants but the amount of prevoicing is not known to vary with the place-of-articulation.

2. Japanese

The Japanese language has a contrast that is similar to the Hindi dental versus retroflex contrast in one voicing context: the Japanese /d/ versus the Japanese /ɾ/ (Best and Strange, 1992).[2] However, there is some uncertainty concerning the articulation of the /ɾ/ in Japanese (Price, 1981). Many linguists describe the Japanese /ɾ/ as a voiced alveolar flap or tap which is sometimes palatalized (Bloch, 1950; Jorden, 1963). Others note that it may be produced as a retroflex (Martin, 1963, "by lifting the tip of the tongue backwards" p. 21; see also Price, 1981). Hattori (1984) and Ogawa (1982) state that in syllable-initial position, the Japanese /ɾ/ is a lax stop (because the aspiration is weaker than for /d/) and that the general place-of-articulation is postalveolar but that the actual place-of-articulation depends on the following vowel (/i/ causes the most forward articulation and /o/ causes the most backward articulation).

More important, the Japanese flap /ɾ/ contrasts with /d/ and /t/ in many environments (but not with any lateral or retroflex phones). Thus, Japanese dental /d/ and alveolar flap /ɾ/ could be considered a phonemic distinction similar to the Hindi dental versus retroflex distinction, probably most similar to the voiced-unaspirated consonants of Hindi. In any case, no study has as yet examined the perception of the Hindi distinction by native speakers of Japanese and pilot work in our laboratory suggests that the dental/retroflex comparison is appropriate.

3. American English

The dental versus retroflex contrast does not occur phonemically in English (Jones, 1964). English does, however, employ alveolar stops which contrast with bilabial and velar stops. Interestingly, dental and retroflex stop consonants occur in English phonetically as allophonic variations of alveolar consonants. And, of course, aspirated and nonaspirated stop consonants occur regularly in English although aspiration is not distinctive (Delattre, 1966). As Polka (1991) pointed out, English /t/ and /d/ in syllable-initial position are typically produced as voiceless-aspirated and devoiced-unaspirated alveolar stops, but they usually are produced as retroflex stops when produced in a consonant cluster with /ɹ/ (e.g., "try" and "dry"). Also, English /d/, in syllable-medial position, may be produced as a prevoiced alveolar stop or as a prevoiced retroflex stop when preceding /ɹ/ (e.g., "added" versus "address" or "night rate" versus "nitrate"). While American English and Japanese both contain an alveolar flap, which may be articulatorilly and acoustically similar in both languages, the flap operates in very different phonological systems. As noted by Price (1981), the English alveolar flap is not generally considered to contrast with /d/ or /t/, particularly in syllable-initial position. In intervocalic position, the English flap contrasts with /l/ and /ɹ/ but not with /d/ or /t/.

Finally, dental /t/ and /d/ occur in final consonant clusters such as "eighth" and "width." And, American English contains the rhotic approximant /ɹ/, which frequently is produced as a retroflex. Thus, American speakers have allophonic experience with dental and retroflex consonants, but only in the presence of certain other phonemes, and not in syllable-initial prevocalic contexts

B. Pretest: Initial status

The Pretest examined the relative difficulty in perceiving the four Hindi dental/retroflex contrasts for three different samples: native American English, native Japanese, and native Hindi speakers. Multiple tokens of Hindi dental and retroflex stop consonants were recorded and presented to participants from these three language groups, and their responses to the various tokens were collected. The tokens were produced by two male native Hindi speakers in four of the five voicing/aspiration contexts with three different vowels (/a/, /e/, /o/) in consonant plus vowel (CV) syllables.

From the description of the phonemic and phonetic make-up of Hindi, American English, and Japanese with regard to dental and retroflex consonants, two alternative outcomes might be predicted. First, it could be argued that American speakers would perform better than Japanese speakers because of their extensive allophonic experience with retroflexion. Alternatively, it might be hypothesized that Japanese participants would perform better than American participants because of their phonemic experience with the contrast between /d/ and /ɾ/ (in spite of their lack of experience with retroflexed /d/ and /t/). A further prediction would be that Japanese speakers would perform best on the voiced-unaspirated dental versus retroflex contrast because it is most similar to their /d/ vs /ɾ/ contrast.

1. Methods

a. Participants. Twenty native speakers of American English, 20 native speakers of Japanese, and 10 native speakers of Hindi were tested. There were equal numbers of males and females in both the American and Japanese groups. The Hindi group had 6 males and 4 females. American and Japanese participants were monolingual and all were free of any hearing deficiencies by self-report (in addition, Japanese participants were given a quick auditory screening threshold test). Although many of the Japanese participants had studied English for several years, few had any extended exposure to speech by native English speakers. None considered themselves fluent in English. Hindi participants were all at least trilingual, with Hindi and English being the common two languages. (However, one female participant in the Hindi group did not actually speak Hindi but did speak Malay, Kannada, and Tamil which are all phonemically similar to Hindi.) American participants ranged in age from 19 to 43 years with a mean of 26 years. Japanese participants ranged in age from 20 to 33 years with a mean of 25 years. Hindi participants ranged in age from 22 to 32 years with a mean of 27 years.

b. Stimuli. All stimuli were produced by two male, native speakers of Hindi. Both were born and raised in India and learned Hindi as their first language. The parents of the speakers were native Hindi speakers and spoke mostly Hindi at home. Both speakers began learning English as a second language between the ages 4 and 6 years when formal education began. Speaker 1 was a 26-years-old male who had lived in the United States for two and a half years. Besides English and Hindi, he also spoke Punjabi. Speaker 2 was a

23-years-old male who had lived in the United States for three years and also spoke Punjabi and Urdu.

The stimuli were recorded on Ampex 631—$\frac{1}{4}$ in. audio tape at $7\frac{1}{2}$ in./s using a Tascam 22-2 two-track reel-to-reel tape recorder via a Shure Spher-O-Dyne (533SA) high impedance microphone. The stimuli were low-pass filtered at 9.7 kHz and converted to digital wave form files at a 20-kHz sampling rate with 12-bit resolution using a pulse code modulation (PCM) hardware coding device (see Whalen *et al.*, 1990) attached to a VAX station computer. The digitized stimuli were pre-emphasized at 12 dB per octave beginning at 800 Hz.

The speakers recorded a random arrangement of consonant-vowel (CV) syllables which were written in standard Hindi orthography. At least 20 exemplars (10 dental and 10 retroflex) were produced by each speaker in each of the four voicing/aspiration contexts with each of the three vowels. Four native Hindi speakers were asked to identify a random presentation of all of the recorded tokens, blocked by speaker and vowel context. Any token that was misidentified by any listener was discarded (75 out of 480 tokens). Finally, for each voicing/aspiration context in each vowel context, eight tokens per speaker (four dental and four retroflex) were selected for the test. The 192 tokens (96 tokens per speaker) were chosen to be similar in amplitude, intonation, and duration. This selection process ensured that nonphonemic variables (e.g., amplitude, intonation, duration) did not vary consistently in relation to phonemic variables (e.g., burst amplitude, VOT, aspiration amplitude). The selected tokens were recorded from the VAX station to a Teac DAT recorder (DA-P20) and then re-digitized to a Macintosh computer at a rate of 44.1 kHz with a 16-bit amplitude quantization. Due to this process, the tokens retained the characteristics of their original digitization format of 20 kHz sampling rate and 12 bit quantization.

c. Procedure. Each participant completed a consent form and a language background questionnaire. Participants were tested individually on a Macintosh computer via headphones (STAX SR Lamda Professional) in a sound attenuated chamber. Prior to testing, each listener was given a brief explanation of the articulatory characteristics of the Hindi dental and retroflex consonants using a diagram of the human vocal tract. Then a complete description of the testing procedure was given. Instructions were given in Japanese for Japanese participants and in English for Hindi and American participants.

Four tests were given; one for each voicing/aspiration context. On each test the speaker and vowel context were randomly ordered. Prior to the administration of each test, a familiarization sequence was presented to the participants in which all 48 stimuli were blocked by speaker, vowel, and place-of-articulation. During this familiarization, participants were instructed to listen carefully without responding in order to become familiar with the particular stimuli for that test. During the familiarization, the computer indicated the speaker, the vowel context, and the place-of-articulation for each stimulus. During the test, the participants heard a single token twice, decided whether the token contained a dental or a retroflex consonant, and then clicked on the chosen response using the mouse. No feedback was given during a test, but after each test, participants were paid 1 cent (or 1 yen) per correct trial in an attempt to provide extrinsic motivation to perform well.

The presentation order of the four voicing/aspiration tests was counterbalanced across participants using a Latin square design that controlled immediate sequencing effects as well as order effects. The entire test lasted approximately $1-1\frac{1}{2}$ h.

d. Analyses. To investigate the relative difficulty of the contrasts in different voicing/aspiration classes, vowel contexts, and speakers for each language group, the percent correct scores were averaged across participants for each context condition.[3] The resulting design was a $3\times4\times3\times2$ repeated-measures model (i.e., 3 language groups $\times$ 4 voicing contexts $\times$ 3 vowel contexts $\times$ 2 speakers) analysis of variance (ANOVA). However, the comparisons of interest were between the American and Japanese language groups. Thus, after this initial analysis, the Hindi data set was removed and a second analysis was done with only the American and Japanese groups. To support this decision, a separate analysis was done on the Hindi data alone to ensure that no significant differences existed between the various stimulus conditions.

2. Results

Correct identifications, averaged across speakers, vowels, and voicing/aspiration contexts, showed that native Hindi participants identified the dental and retroflex stimuli more accurately (96% correct), than native Japanese participants (71% correct), who in turn were more accurate than native American participants (59% correct); $F(2,47)=87.97$, $p=0.0001$. Subsequent testing revealed that all three means were reliably different in pairwise comparisons. The near perfect performance of the native Hindi participants indicates that the stimuli used in the test were correctly produced and that the task did not tax the native participants' general perceptual or cognitive abilities. Further, as expected, the ANOVA on Hindi participants' data alone showed no reliable differences among the various stimulus conditions (no speaker effect, no vowel context effect, and no voicing/aspiration differences).

All three language groups were more accurate in classifying productions by Speaker 1 than those by Speaker 2 [78% vs 73% correct; $F(1,47)=35.13$, $p=0.0001$]. This difference in accuracy between the speakers was greatest for American participants [a difference of 9.1%; $F(1,19)=38.57$, $p=0.0001$], slightly less for Japanese participants [a difference of 5.5%; $F(1,19)=13.38$, $p=0.0017$], and trivial for Hindi participants [a difference of 1.8%; $F(1,9)=2.32$, $p=0.162$].

Perceptual differences as a function of vowel context were not as systematic across language groups. A distinct order of difficulty was found for Japanese participants [from most difficult to least difficult, /a/ 67% correct, /e/ 70%, and /o/ 77%; $F(2,38)=38.81$, $p=0.0001$]. No reliable differences in vowel context were found for either the American or Hindi groups.

Finally, perceptual differences as a function of voicing/aspiration context were somewhat systematic across language groups. While no differences existed for Hindi participants [$F(3,27)=1.14$, $p=0.3523$], reliable differences were observed for both American and Japanese participants

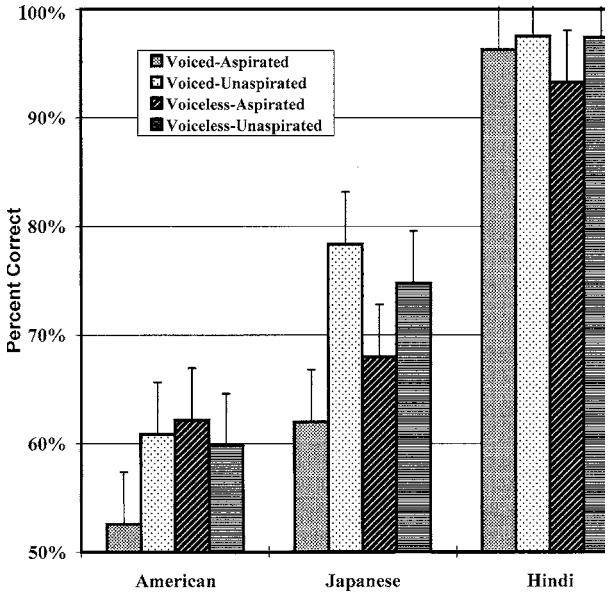


FIG. 1. Accuracy of identification of the Hindi dental-retroflex distinction in four different voicing/aspiration contexts by American, Japanese, and Hindi participants. The error bars represent Fisher's least significant difference values.

[respectively, $F(3,57)=7.15, p=0.0004$, and $F(3,57)=12.89, p=0.0001$]. As illustrated in Fig. 1, American participants performed more poorly on the voiced-aspirated context than the other three contexts. Japanese participants also performed most poorly on the voiced-aspirated context, and, more generally, performed significantly worse on the two aspirated contexts than on the two unaspirated contexts.

3. Discussion

Several results are of interest in this study. First, a clear difference emerged between the two groups of non-native participants. Japanese participants showed a distinct advantage over the American participants in identifying the Hindi contrast. This advantage was found in each of the phonetic contexts and speaker conditions. In particular, while Japanese participants performed best on the pair that is most similar to their native language phonemes (voiced-unaspirated dental versus retroflex), they also outperformed American participants on the other voicing/aspiration contexts. This finding supports the hypothesis that phonemic (contrastive) experience with a place-of-articulation contrast similar to the Hindi dental-retroflex contrast leads to better performance on that non-native contrast than experience which is primarily non-contrastive even though allophonically rich.

Best has developed an assimilation model that can help to explain these data (Best, 1993, 1994). She argues that listeners must categorize the target speech sounds (and possibly provide a "goodness of fit" value) according to their own native language speech categories. She has predicted varying degrees of perceptual difficulty based on the assimilation pattern that emerges. For example, when two non-native sounds fit nicely into two different native categories, the contrast is perceptually easy to make. On the other hand, if the two non-native sounds fit into only one native category, the contrast is more difficult perceptually. Several other patterns are possible; however, in the present study, these two patterns are what emerged. From informal discussions with participants following the pretest, many Japanese participants assimilated the Hindi dental stop to their Japanese dental stop /d/ and they assimilated the Hindi retroflex stop to their flapped /r/. American participants tended to categorize both Hindi consonants as the alveolar /t/ or /d/ or the dental /θ/ or /ð/ depending on the voicing/aspiration context. Occasionally, they distinguished the two as if they were a voicing contrast /t−d/ or /θ−ð/.

Vowel context only affected the performance of Japanese participants; American participants performed equally poorly in all three vowel contexts. These vowel effects may have occurred because of coarticulation effects for both Hindi retroflex consonants and the flapped /r/ of Japanese, as noted previously (Dave, 1977; Dixit, 1990; Dixit and Flege, 1991; Hattori, 1984; and Ogawa, 1982), or may be a "floor effect" for the American participants, given their overall poor performance and a "ceiling effect" for the Hindi participants given their superior performance.

III. TRAINING ON THE HINDI DENTAL-RETROFLEX CONTRAST

The pretest showed that Japanese participants were markedly better than American participants at perceiving the Hindi contrast prior to any training. Given this initial difference, a training study, complete with multiple generalization tests, was conducted to further elucidate the role of native language experience in facilitating or interfering with learning to perceive a difficult non-native contrast. The participants who had been pretested in Study 1 were given 12 short but intense training sessions which included a subset of the materials from the pretest. Following this training, the identical test was given as a posttest and performance on both trained and nontrained items was assessed.

A. Background

Over the past three decades there have been a number of experimental training studies aimed at enhancing the perception of foreign speech sounds. The favorite targets have been: (1) improving the perception of a new distinction in voice-onset-time in initial stop consonants (e.g. Carney *et al.* 1972; Edman, 1980; Jamieson and Morosan, 1986, 1989; Lisker and Abramson 1970; McClaskey *et al.*, 1983; Moroson and Jamieson, 1989; Pisoni *et al.*, 1982; Strange, 1972), (2) improving the perception of the American /r/−/l/ distinction by Japanese listeners (e.g., Bradlow *et al.*, 1999, 1997; Flege *et al.*, 1996; Hardison, 2003; Lively *et al.* 1993, 1994; Logan *et al.*, 1991; McCandless *et al.*, 2002; Strange and Dittmann, 1984; Yamada, 1993, 1995), and (3) improving the identification of the Hindi dental/retroflex contrast by American listeners (Pruitt, 1992; Pruitt *et al.*, 1990; and Werker and Tees, 1984). One generalization that seems warranted by this collection of studies is that voicing distinctions seem to be easier to train than place-of-articulation distinctions [see, especially, Edman (1980) and Werker and Tees (1984) for specific contrasts]. This may be attributable to the

relative independence of the motor control of voicing from the control of other articulators and the related perceptual dependence on timing differences rather than spectral differences.

Bradlow (in press) has recently surveyed the /r/−/l/ training studies to summarize what they have taught us about the training of cross-language distinctions. She drew three major generalizations:

(1) The ability to modify speech perception and production patterns is retained in adulthood. Laboratory-based training studies can lead to increased accuracy in the perception of non-native contrasts even for difficult instances such as /r/−/l/.
(2) Learners can acquire better functional category representations, although it is not known whether they employ the same acoustic features as those the native speakers use. Bradlow recommends that trainers should be especially careful to train contextualized speech samples up to and including sentences and larger discourse units.
(3) Highly variable training stimuli promote, rather than interfere with, adequate acquisition of non-native contrasts. Such variability should include multiple talkers, multiple phonetic contexts, and multiple syllable positions of the contrast.

Preliminary studies in our laboratory focused on the problems of training American English monolinguals to identify the Hindi dental/retroflex contrast (e.g., Polka, 1989, 1991; Pruitt *et al.*, 1990; Pruitt, 1992; Strange *et al.*, 1989). These studies, and the others given earlier, provided several findings that guided the present study. First, there were reliable differences in trainee's performance prior to and after training that depended on the vowel context, voicing/aspiration context, and the speaker used in training. Second, little of what was learned during brief training transferred to nontrained stimuli that differed in vowel context, voicing/aspiration context, or speaker. Third, truncation of the vowel portion of consonant-vowel syllables increased the discriminability of the consonants and accelerated progress in training, presumably by isolating the important acoustic cues. Fourth, training with truncated stimuli alone did not improve posttest performance with full-length (unaltered) stimuli any more than full-syllable training. Finally, interactive self-paced training that permitted some selection of materials appeared to produce more learning than traditional trial-oriented training methods.

Review of the training studies argues that tests of transfer of training (i.e., generalization of learning) are important for the evaluation of training success. Generalization helps identify what the learner has actually learned. However, transfer of training is not a simple unidimensional construct. Transfer of training in speech perception can be divided into at least seven categories: (1) transfer to new tasks (e.g., from discrimination to identification with the same stimuli), (2) transfer to new instances from the same speaker (i.e., same context, same contrast, but new tokens), (3) transfer to new speakers, (4) transfer to new phonetic contexts (e.g., from consonant plus /e/ to consonant plus /o/), (5) transfer to new syllabic environments (e.g., from an initial consonant cluster to a final consonant cluster), (6) transfer to new contrasts of the same feature (e.g., dental versus retroflex in voiced stop consonants to dental versus retroflex in voiceless stop consonants), and (7) transfer from isolated tokens to tokens produced in context (i.e., in running speech). See Logan *et al.*, 1993; Logan and Pruitt, 1995, for a full discussion. These types of transfer differ in the amount and type of nonrelevant variation that the listener must ignore to make a correct response.

Preliminary studies mentioned earlier demonstrated that native English listeners could improve their identification of tokens of the Hindi dental/retroflex contrast with only brief training. However, learning was slow and effortful, and gains were modest (10%–16%). In addition, the improvement produced by training did not transfer to new (nontrained on) stimuli. For the present study, several modifications were made in the training program. First, the number of training sessions was increased to 12. Second, several of Skinner's (1954, 1958) concepts for "automatic teaching machines" were integrated into the new perceptual training program. Specifically, the new training procedure made use of gradual, "small-step" progression of the difficulty of the training materials, and made presentation of successive items conditional in part upon previous performance. The training stimuli were modified, using vowel truncation, to permit training to progress in a gradient fashion from the easiest distinctions (truncated stimuli) at the beginning of training to the more difficult distinctions (full syllables) at the end of training. It has been argued that gradient training is more effective than fixed-difficulty training (Jamieson and Morosan, 1989; McCandless *et al.*, 2002), and, as our own research has shown, truncation of the vowel portions of CV syllables increases the discriminability of the consonants.

A final change in the training procedure was that the training stimuli were produced by six speakers thus increasing the acoustic variability within a phonetic category (an important factor noted by Logan *et al.*, 1991). This variability is believed to help the participants abstract more general aspects of the dental-retroflex distinction rather than allowing them to concentrate on a few acoustic idiosyncrasies that might be present in a small set of stimuli from a single speaker. During training, new speakers were added in a gradient fashion along with the truncation gradient.

B. General design

A pretest-training-posttest design was employed to assess participants' initial ability (Study 1) and the effects of training on perception of the Hindi contrast. The 20 native English and 20 native Japanese participants from Study 1 were given training and then retested. The posttest was identical to the pretest.[4]

Training was conducted using only a subset of the pretest stimuli, in addition to many new stimuli described in the following. Tokens from one of the two pretest-posttest speakers were used in training and tokens from the other were not. The particular training speaker employed was counterbalanced across the participants. The amount and type of learning was assessed by multiple tests of transfer-of-

training. Two of the vowel contexts from the pretest-posttest (/a/ and /o/) were used in training and one context was not. One of the voicing/aspiration contexts from the pretest-posttest (breathy-voiced) was used in training and three others were not. Thus, three types of transfer could be assessed at posttest: (1) transfer to a speaker who had not been included in training, (2) transfer to a different vowel context, and (3) transfer to three different voicing/aspiration contexts. The voicing context chosen for use in training was the voiced-aspirated context that was the most difficult for both American and Japanese participants in the Pretest. Thus, it offered the greatest challenge and the opportunity for the greatest gain.

1. Methods

a. Training stimuli. In addition to the two male speakers from the pretest-posttest, five additional speakers (two males and three females) were recorded. The five additional speakers were all born and raised in India and learned Hindi as their first language except for one of the females who learned Konkani first. All were multilingual and spoke at least two languages (Hindi and English). All Hindi speakers began learning English as a second language between 4 and 6 years of age when formal education began. The speakers ranged from 24 to 30 years old and had lived in the United States between 2 and 5 years. During training, one of the two speakers from the pretest was also used, thus, each trainee heard a total of six training speakers. The recording procedure was identical to that used in Study 1 except that the stimuli were recorded through an AKG C1000s microphone directly to DAT and then digitized from DAT to a Macintosh at a 44.1-kHz sampling rate with 16-bit quantization. Thus, these additional training stimuli were somewhat higher fidelity than the pretest-posttest stimuli.

The training stimuli were edited to create the vowel-truncated stimuli. Four different levels of truncation were created. The most severely truncated stimuli consisted of the burst (plus any prevoicing) and the subsequent 60 ms (which was primarily aspiration). The next level of truncation included an additional 40 ms (i.e., 100 ms plus any prevoicing). The third level included an additional 60 ms (i.e., 160 ms plus any prevoicing). The final level was not truncated, that is, the entire token was included. These amounts of vowel truncation were chosen through earlier pilot testing and subjective evaluation. In truncating the stimuli, a 20-ms window preceding the terminal cut was selected and the amplitude of the stimulus in this window was tapered to zero to prevent "pops" or "clicks."

b. Training procedure. The participants in each language group were randomly assigned to one of two training conditions. Half the participants trained with Speaker 1 from the pretest as their "base speaker" and the other half trained with Speaker 2 as their base speaker. Participants in the first condition always had Speaker 1 as a listening choice and always had at least one test item from that speaker on every mini-test from the very beginning of training. Participants in the other training condition had Speaker 2 in similar fashion.

Participants received 12 training sessions administered on an individual basis. Up to three training sessions could be completed in one day if the listener's schedule required it. Training took place in a sound attenuated chamber on a Macintosh computer via STAX SR Lamda headphones using a HyperCard program that controlled the presentation of the stimuli and collected the participants' performance data. Participants were allowed to adjust the sound volume to a comfortable listening level. Training proceeded as a series of alternating listening and testing blocks. During a listening block, by selecting specific icons, participants could select the speaker, vowel context, and place-of-articulation of the tokens for listening. The participants were instructed to try to learn the difference between dental and retroflex consonants by listening to these consonants presented in the various possible contexts in any way that they wished.

There were ten listening blocks and ten mini-tests in each training session. A listening block included 50 presentations of the tokens selected by the participant. (This procedure is similar in some respects to the category-switch task of Tees and Werker, 1984.) A mini-test was a quasirandom selection of ten tokens from the training materials of that session. Randomization was subject to the following constraints: Every speaker available in the session was represented at least once; equal numbers of dental and retroflex tokens were used; and, the same type of token was not allowed to succeed itself more than twice. Tokens were selected without replacement. Tokens were presented twice in a trial (as in the pretest-posttest) with immediate feedback following the response. When a listener responded incorrectly to an item on the mini-test, the token was presented again with the correct response. (This was similar to the technique of Logan *et al.*, 1991.) Participants were given small monetary rewards for correct performance on the mini-tests (two cents or two yen for each correct trial, a maximum possible bonus of $2.00 per session). Previous research in our laboratory suggested that this incentive was required because participants had little intrinsic motivation to learn the Hindi consonants.

The gradient technique using truncated stimuli and the number of speakers proceeded as follows. On the first day of training, participants started with only one speaker (the "base speaker"). All tokens were the most severely truncated type. If the listener scored 9 or 10 out of 10 items on a mini-test, an additional speaker was made available for listening. Once all six speakers were available for listening, continued good performance (9 out of 10 or better) led to the introduction of stimuli with less vowel truncation. Regardless of the listener's performance at the end of a training session, each new training session started at an easier level. However, even if a listener had not progressed during a particular training session, an additional speaker was made available at the beginning of every odd-numbered training session. This procedure ensured that by session 11, every listener had gained experience with all six of the speakers. In addition, on training session 7, the second level of truncation became the new starting point. These rules for changing the gradients forced the participants who were performing less well to experience the more difficult materials while allowing the more advanced participants to reach the more difficult levels sooner. This training method exposed participants to the natural variability of the dental-retroflex contrast in the different contexts described earlier.

At least 700 tokens were presented to the participant during each session of training, giving a minimum of 8400 tokens in total by the end of the 12th training session. Each session lasted between 30 and 50 min. A questionnaire was given following the final training session that asked participants about their experience with the training. One part of

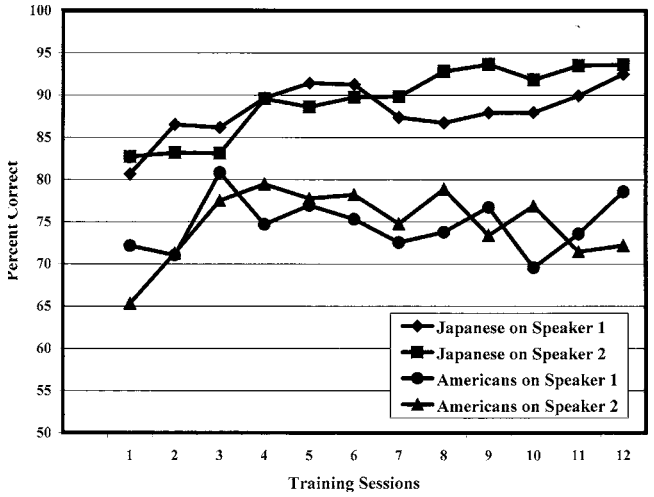


FIG. 2. Percent correct identification on the "Basic Training Speaker" tokens in the training mini-tests for each training session, plotted by language group and speaker.

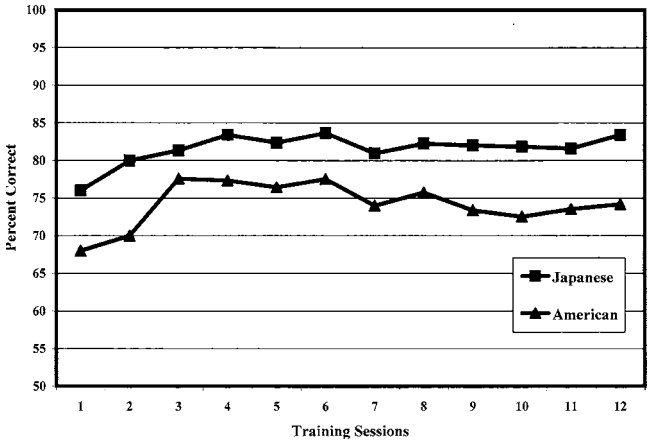


FIG. 3. Percent correct identification on all mini-tests in each training session, plotted by language group.

the questionnaire asked participants to simply write down what sounds the Hindi dental and retroflex consonants sounded like in their own language.

2. Progress through training

Two measures of training performance were examined. The first measure was simply the percent correct score per training session on the tokens from the "base speaker" alone. It will be recalled that there are some data for the base speaker for every mini-test for every participant; thus, such a measure provides an estimate for each group as to their performance on this most-practiced individual. It should be remembered, however, that the individual tokens even from this speaker became more difficult as syllable truncation was graduated during training.

Figure 2 shows the Base Training Speaker plots for the Japanese and American trainees. To the extent that gradient training is appropriately applied, it should produce fairly flat performance curves very similar to those shown in Fig. 2. There is little effect of particular speaker but a marked difference between the performance levels of the two language groups. These data suggest that the rate of graduation of difficulty was about right for the American participants but that the increase in difficulty could have been more rapid for the Japanese group.

A second performance measure was calculated by averaging the percent correct performance in each language group on the ten mini-tests in each session. Strictly speaking, the numbers are not comparable from trial to trial or participant to participant. The particular selection of training items, the number of different speakers available to each listener, and the amount of truncation of the stimuli differed from participant to participant and from trial to trial. Thus, this measure, too, is only an approximate index of progress of the language groups. The result is shown in Fig. 3 where it can be seen that there is a consistent difference between the language groups in favor of the Japanese participants. It appears that there is some adaptation to the task over the initial sessions for both groups but then performance is maintained at a level rate of success over the course of the sessions.

Not revealed in either of the training session graphs is the fact that many Japanese participants performed perfectly (i.e., 100% correct) on certain speakers, especially during the last several training sessions. This was rarely the case with American participants.

a. Posttest and generalization. Upon completion of the training sessions, each participant was given a posttest that was identical to the pretest. Because training was conducted on only a subset of the pretest materials (one voicing/aspiration class, two vowel contexts, and one base speaker), it was possible to measure generalization to the three other voicing/aspiration classes, to the other vowel context, and to the other speaker whom they had not heard since the pretest.

3. Results

a. Analyses. A mixed design analysis of variance (ANOVA) was conducted on pretest and posttest percent correct scores of trained-on materials (i.e., voiced-aspirated, /a/ and /o/ syllables) to determine whether training had a differential effect related to native language. Similar analyses were also conducted on the material that had not been used in training in order to assess transfer of training for each language group.

b. Pretest equivalence. Participants in both the Japanese and American groups had been randomly placed into one of two conditions: training with Speaker 1 or Speaker 2 as the base speaker. An ANOVA was calculated to determine whether the participants in these two groups differed from each other on the Pretest. No differences were found except for a three-way interaction among the voicing/aspiration context, the Pretest Speaker, and the Base Speaker. This three-way interaction was neither patently systematic nor inherently interesting and will not be discussed further. The main finding from this analysis was that the trainees in the two speaker-training groups, pooled over Japanese and English trainees, did not differ reliably at pretest. Differences between the speaker-training groups found at posttest may thus be attributed to their respective training conditions and language groups rather than initial differences in accuracy.

c. Posttest results for trained-on materials. An ANOVA was calculated on the percent correct scores for just the training stimuli that were included in the pretest and

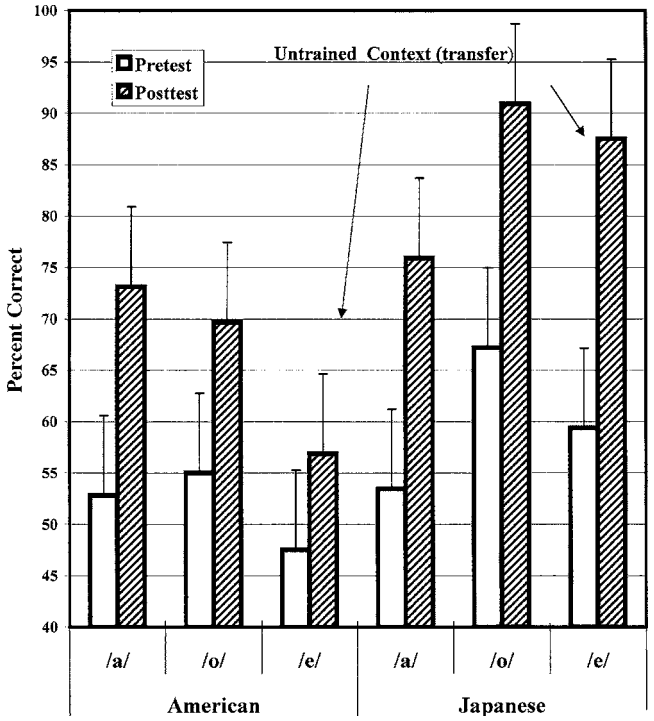


FIG. 4. Percent correct identification on trained and untrained vowel contexts in pretest and posttest, plotted by vowel and language group. (Transfer across vowels.) The error bars represent Fisher's least significant difference values.

posttest (that is, voiced-aspirated stimuli produced in the /a/ and /o/ vowel context). This ANOVA included the between-subjects variables, Language Group (American versus Japanese) and Training Speaker (Speaker 1 versus Speaker 2), as well as the within-subjects variables, Vowel Context (/a/ vs /o/) and Pre-Post (Pretest versus Posttest).

As expected, there was a reliable increase in overall performance from pretest to posttest on the trained-on materials [from 57% to 77%, an increase of 20 percentage points in accuracy; $F\ (1,36)=94.56$ $p=0.0001$]. While Japanese participants performed better than American participants, collapsed across the pretest and posttest [72% vs 63%, a difference of 9%, $F\ (1,36)=8.95$, $p=0.005$], the interaction of Language Group with Pre-Post score was not significant; that is, the gains of the two groups were not reliably different. The mean pretest to posttest percent correct score for Japanese participants on the training items was 60%–83%, a gain of 23 percentage points, and for Americans it was 54%–72%, a gain of 18 percentage points. However when the data were collapsed across the pretest and posttest, there was an interaction between Language Group and Training Speaker. Japanese and American participants performed similarly to each other on Speaker 1 tokens, 69% vs 67%, respectively, but differently on Speaker 2 tokens, 75% vs 58%, respectively [$(F(1,36)=5.44)$, $p=0.0254$]. Additionally, and again at both pretest and posttest, the vowel context /a/ was more difficult than /o/ [64% vs 71% correct; $F\ (1,36)=10.77$, $p=0.023$]. This vowel context effect interacted with the Language Group [$F\ (1,36)=12.81$, $p=0.001$]. As shown in Fig. 4, Japanese participants had more difficulty at both pretest and posttest with the /a/ vowel context than the /o/ vowel context while American participants performed similarly on the two vowels. This effect was similar to the Pretest results described earlier. This finding indicates that some differences in perceptual difficulty may be stable even after intensive laboratory training.

Finally, there was differential improvement from pretest to posttest for participants of both languages depending on vowel context and training speaker [$F\ (1,36)=4.16$, $p=0.0489$]. Participants trained with Speaker 2 improved less on the /o/ vowel context than on the /a/ context, but those trained on Speaker 1 showed the most improvement for the /o/ context. There was no statistically reliable interaction of Language, Vowel Context, and Pre-Post, or of Language, Vowel Context, Pre-Post, and Training Speaker.

d. Posttest results (transfer of training). Turning now to transfer of training (i.e., pretest to posttest change on trained versus untrained tokens and trained versus untrained contexts), only the incidents in which there was a significant interaction of an independent variable with the Pre-Post variable will be reported. This interaction indicates a differential change from pretest to posttest. Lack of such an interaction (in the presence of reliable change overall from pretest to posttest) means that the training successfully transferred to the nontrained materials. In other words, in those cases participants improved similarly on trained and nontrained tokens.

e. Transfer to a new vowel context. Transfer of training to a new vowel context, /e/, is also shown in Fig. 4. Transfer was assessed by conducting an ANOVA on participants' performance with the voiced-aspirated, /a/ and /o/ tokens (i.e., training stimuli) versus the untrained-on voiced-aspirated /e/ vowel context tokens. Language Group and Training Speaker were included as between-subjects variables and Pre-Post and Vowel Context (/a/, /e/, /o/) were included as within-subjects variables.

The analysis found no differential amount of improvement from pretest to posttest for any vowel context; that is, transfer of training occurred to the untrained-on vowel context /e/. Statistical contrasts confirmed that reliable changes from pretest to posttest occurred in all three vowel contexts for both Language groups. Collapsed across all three vowel contexts, participants reliably improved in performance from pretest to posttest; from 56% to 76%, a gain of 20 percentage points [$F\ (1,36)=85.64$, $p=0.0001$]. Additionally, there was an interaction between Pre-Post and Language Group [$F\ (1,36)=5.47$, $p=0.0251$]. American participants improved less from pretest to posttest, 52%–67%, a gain of 15 points, than did Japanese participants, 60%–85%, a gain of 25 points.

Collapsing across the pretest and posttest, there was a reliable effect of Vowel Context, a Vowel by Language Group interaction, and a Vowel by Training Speaker interaction. These effects did not interact with Pre-Post and thus are not different from those already described in the results of the Pretest and will not be discussed further.

f. Transfer to a new speaker. Transfer of training to a new speaker was assessed by conducting an ANOVA on participants' performance with voiced-aspirated, /a/ and /o/ tokens (i.e., trained contexts) for familiar versus unfamiliar speakers. This analysis included Language Group and Training Speaker as between-subjects variables, and, Pre-Post, Speaker Familiarity, and Vowel Context as within-subjects variables. Overall, participants improved reliably in performance from pretest to posttest on both speakers; 58%–76%, a gain of 18 percentage points [$F\ (1,36)=196.33$, $p=0.0001$]. However, there was a signifi-

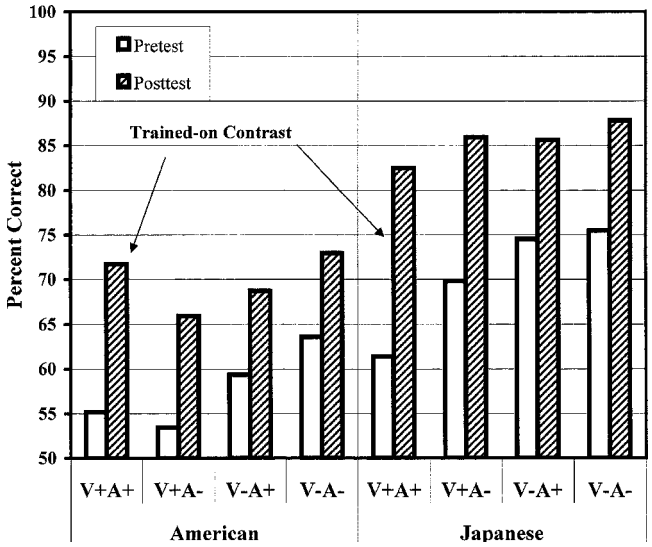


FIG. 5. Percent correct identification for trained and nontrained voicing/aspiration contexts in pretest and posttest, plotted by language group and voicing/aspiration context. (Transfer across voicing/aspiration contexts).

cant interaction of Pre-Post and Language Group. American participants improved less than Japanese participants from pretest to posttest: American, 53%–67s%, a gain of 14 percentage points versus Japanese, 62%–83%, a gain of 21 percentage points [F (1,36), $p=0.0447$]. In addition, there was a significant interaction of Pre-Post with Speaker Familiarity and Training Speaker [F (1,36)$=6.38$, $p=0.0161$]. Subsequent comparisons showed that while performance on both speakers was about equal at pretest, participants in the Speaker 1 condition showed much better performance on Speaker 1 than on Speaker 2 at posttest. However, participants in the Speaker 2 condition performed about equally on the two speakers at posttest. However, there was sufficient improvement from pretest to posttest on the untrained-on speaker condition to indicate that transfer of training did occur for both language groups and both training speaker conditions.

g. Transfer to new voicing/aspiration context. Transfer to unfamiliar voicing contexts was assessed by conducting an ANOVA on participants' performance with /a/ and /o/ tokens (i.e., training contexts) for the familiar voiced-aspirated context versus the three untrained-on voicing/aspiration contexts (voiced-unaspirated, voiceless-aspirated and voiceless-unaspirated). Once again, Language Group and Training Speaker were included as between-subjects variables and Pre-Post, Vowel Context (/a/ and /o/), and Voicing/aspiration Context were included as within-subjects variables.

There was a reliable interaction of Pre-Post with Voicing/aspiration Context [F (3,108)$=3.46$, $p=0.019$]. Figure 5 shows that, independent of Language Group, participants improved more from pretest to posttest on the trained-on voicing/aspiration context (voiced-aspirated) than on the other three contexts. Still, as evident in Fig. 5, there was significant improvement for each voicing/aspiration context indicating that transfer of training did occur to all three nontrained voicing/aspiration contexts. This effect was also influenced by the Training Speaker and the particular voicing/aspiration Context [F (3,108)$=4.01$, $p=0.0095$]. Participants trained with Speaker 2, pooled across the Language Groups, did not improve reliably on the voiceless-unaspirated context (70% vs 74%).

h. Assimilation data. The post-training questionnaires revealed that during training most American participants attempted to differentiate the voiced-aspirated dental and retroflex consonants (i.e., the training stimuli) as a voicing contrast while Japanese participants attempted to differentiate these consonants as a place-of-articulation contrast. Table I shows the frequency of occurrence of several native language phoneme labels for each language group. (Only consonant labels given by more than one listener are shown in Table I.)

As can be seen, American participants primarily assimilated dental consonants to "t," "d," and "th" (/θ or ð /) and retroflex consonants consistently to "d." Japanese participants, on the other hand, assimilated dental stops to Japanese "t" and "d" and retroflex stops to Japanese "r." Note that the data in Table I do not reveal how many participants actually categorized dental versus retroflex as two different consonants. A counting of such cases (defined as when a listener did not use a label twice—once for dental and once for retroflex) showed that 11 American and 13 Japanese participants had a clear two-category assimilation pattern. Of the participants who reported such a pattern, all 11 of the American and one of the Japanese reported the assimilation as "t" versus "d," while none of the American and 12 of the Japanese reported the assimilation as "t" or "d" versus "r."

IV. DISCUSSION

The aim of the training study was to determine whether the more accurate performance of Japanese listeners over American participants at pretest would produce differences in learning and generalization after intensive laboratory training. In the present study Japanese participants classified tokens more accurately than American participants on Pretest, during training, and on Posttest. This superiority extended both to trained-on and to transfer materials.

Both language groups improved markedly from Pretest to Posttest on the trained-on stimuli, an average improvement of about 20 percentage points. This result is a substantial improvement over previous training studies of Americans on these Hindi contrasts (about 10%, Pruitt *et al.*, 1990, about 16%, Pruitt, 1992); and other non-native contrasts

TABLE I. Participant's reports of the similarity of the training stimuli (voiced-aspirated dental and retroflex consonants) to speech sounds of their native language.

		"t"	"d"	"r"	"th"	"p"
Americans	Dental consonants	15	8	0	5	2
	Retroflex consonants	3	15	1	1	0
Japanese	Dental consonants	7	14	0	0	0
	Retroflex Consonants	0	3	16	0	0

(e.g., /r/ and /l/ training; 12%, Lively *et al.*, 1994; however, see Yamada, 1993, Bradlow *et al.* 1997,1999 where improvement was about 20% after 45 sessions). Importantly, training transferred to an untrained-on speaker, a new vowel context, and to all three nontrained voicing/aspiration contexts for both Japanese and American participants. This outcome shows a substantial improvement over training techniques used in many previous studies which found little or no transfer of training to nontrained stimuli (e.g., Logan *et al.*, 1991; Pruitt, 1992; Pruitt *et al.*, 1990) although Logan *et al.* (1991) found transfer to new words spoken by the trained-on talker and the above-mentioned Bradlow *et al.* studies found generalization to new words spoken by a new talker.

An important result of the present study is that transfer occurred across multiple dimensions simultaneously. For example, learners improved on voiceless-unaspirated, C+/e/ tokens when they were trained only on voiced-aspirated, /a/ and /o/ tokens produced by different speakers. Thus, learning generalized across both the voicing and aspiration dimensions, to a different vowel context, and to a different speaker all at the same time. This extensive generalization occurred for all three voicing/aspiration contexts and was independent of the speaker used in training.

Given the amount of learning and generalization that occurred, the present findings provide strong evidence against a more traditional position that the capacity to learn new speech contrasts declines irreversibly with age. The study further supports Bradlow's (in press) conclusions that there is considerable capacity to learn new speech contrasts in adulthood and that highly variable training materials can support extensive generalization. More advanced training methods, like the ones used by Hardison (2003), Jamieson and Morosan (1986, 1989), Logan *et al.* (1991), McCandless *et al.* (2002), Rochet (1995), and Yamada (1995), and the methods used in the present study are achieving more success than earlier attempts at training. Current training methods involve more complicated feedback schemes, careful selection of stimuli to provide ample variability, or restrict variability when necessary, and motivational regimes such as monetary rewards.

The results of the current study shed some light on the question raised by Jenkins and Yeni-Komshian (1995) concerning the appropriate levels of analysis in cross-language research. Those writers suggested that the appropriate level might be somewhere around the allophonic level and not at the level of systematic phonology, or the phoneme, or even the level of the phonetic unit. Thus, when predicting which contrasts of a certain language would be perceptually difficult for speakers of another language, they suggested that one must consider the specific phonetic environments of the second language in which the contrast occurs. Adequate training of such a contrast, in their view, might need to include all of the various contexts in which it occurs. However, the fact that generalization was found so abundantly in the present study argues that their view is too restricted. Clearly, participants abstracted some feature or features of the dental-retroflex contrast that generalized across these multiple conditions. The effective level of analysis may be somewhat more abstract than previous studies suggested. It is still likely, however, that the appropriate level will depend upon the particular languages, phoneme contrasts, and the materials employed in training.

The post-training assimilation questionnaires revealed that most American participants attempted to differentiate the voiced-aspirated dental and retroflex consonants as a voicing contrast (e.g., /t/ versus /d/) while some labeled both consonants as /d/. Japanese participants, on the other hand, appeared to differentiate these consonants as a place-of-articulation distinction (dental tokens were heard as the Japanese /t/ or /d/ and retroflex tokens were related to the Japanese flapped alveolar /r/). Best's assimilation model (Best, 1993, 1994) fits these data and helps to explain the differential performance of American and Japanese participants. In the present study, the majority of Japanese participants (13 out of 20) showed a two-category assimilation pattern on the voiced-aspirated training stimuli which, compared to American participants, led to more accurate perception of the non-native Hindi contrast.

Jenkins and Yeni-Komshian (1995) briefly discuss the possibility of a situation in which listeners, because of allophonic variation, sometimes hear a non-native phoneme as one native category and sometimes as another. Such situations may occur more often than previously considered. Assimilation studies in our laboratory with non-native vowels (Levy and Strange, 2002, Strange *et al.*, 2004) have found several cases where one non-native vowel is heard equally well as two different native vowels, depending on the consonant context in which it occurs. This possibility is not specifically considered by Best's theory but could explain differences in performance due to vowel context or voicing/aspiration context found in the present study.

The results of this study suggest that while new phoneme contrasts can be learned, some native language influences are not easily overcome. The differences between American and Japanese participants persisted throughout training and were evident in transfer conditions as a simple main effect: In all respects Japanese participants performed the task more accurately. Transfer was found relatively evenly across the different contexts for both language groups. However, contexts that were initially more difficult, relative to other contexts, remained so.

Limitations and open questions: This study would have profited by the inclusion of specific control groups for the American trainees and by independent studies of perceptual assimilation. Future studies should attend to these limitations, especially if they are directed at problems that have not been explored in many pilot efforts. In the present case, see footnote 4, we do not believe the omission jeopardizes the conclusions of the study.

The research uncovered some issues for future study. The differences associated with Training Speakers point to a somewhat neglected variable, namely, the productions of some speakers may be especially helpful or perhaps, especially misleading for the trainees. It should be noticed here that a speaker difference emerged even in the context of five other training speakers. Hardison (2003) also found a speaker effect. This may have considerable practical importance and deserves further study. Similarly, the phonetic con-

text in which the to-be-trained distinction occurs has an effect on the ease or difficulty of the training: here, the following vowel plays such a role for the Japanese trainees but not for the Americans. Finally, neither this study nor any other experimental study has studied the problem of learning a particular distinction in the midst of the entire matrix of new distinctions required by a new language. For example, as one anonymous reviewer pointed out, Hindi also contains a retroflex tap/flap that might well compete with the distinction trained here. A broad study of the perceptual assimilation of the L2 phonemes to the L1 phonemes prior to any training would be a useful adjunct to the research.

ACKNOWLEDGMENTS

This research has been supported by a Dissertation Research Award to J. S. P. from the American Psychological Association and by NIDCD Grant No. DC00323 to W. S. The authors gratefully acknowledge the substantial support of this research by Reiko Akahane Yamada and the Human Information Processing Research Laboratory, Advanced Telecommunications Research Company, Kyoto, Japan. The research would not have been possible without their extensive support and collaboration. We are also grateful to our Hindi speakers and participants for many hours contributed to this study.

[1]A dental consonant is produced by making a constriction of the vocal tract with the tongue tip at or immediately behind the upper front teeth (i.e., the upper incisors). A retroflex consonant is produced by curling the tip of the tongue upwards towards the hard palate to make a constriction somewhere behind the alveolar arch usually with the underside of the tongue (Ladefoged, 1982).

[2]The Japanese /d/ is produced as a short voiced dental stop (Bloch, 1950; "…with the tongue touching the teeth…," Jorden, 1963, p. xxvi). The Japanese /r/ is produced as a flap ("…with a brief flick against the alveolar ridge…," Martin, 1963, p. 21). According to Martin, there are two primary differences between the Japanese /d/ and /r/: "length--the **r** is brief; the **d** somewhat longer; and position of contact--the **r** is against the alveolar ridge with very tip of the tongue, the **d** is against the teeth with somewhat more of the tongue" (p. 21).

[3]Due to the nature of identification tasks when participants are unfamiliar with the categories to be labeled, listeners sometimes reverse the labels of the categories. This was not considered a perceptual error on the listener's part because the labels were determined somewhat arbitrarily (although the categories were not). Following the procedure of Polka (1989), if a listener scored significantly below chance on one of the voicing context tests, the labels were reversed and the score was transformed for that test only. For example, 39% is significantly below chance on a test with 96 trials, so the score would become 61% correct. This correction was applied only once for a Japanese listener in the voiceless-unaspirated context condition.

[4]Because of our extensive experience attempting to train perception of this distinction with American trainees, no separate American control group was tested. However, a Japanese control group of 20 participants was studied by Pruitt and Akahane-Yamada and mentioned in a report at the Acoustical Society of Japan in 1997. On test-retest the control group showed an overall increase of only 3 percentage points.

Best, C. T. (**1993**). "Emergence of language-specific constraints in perception of non-native speech perception: A window on early phonological development," in *Developmental Neurocognition: Speech And Face Processing In The First Year Of Life*, edited by B. de Boysson-Bardies S. de Schonen, P. Jusczyk, P. MacNeilage, and J. Morton (Kluwer, Dordrecht).

Best, C. T. (**1994**). "The emergence of native-language phonological influences in infants: A perceptual assimilation model," in *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words*, edited by J. G. Goodman and H. C. Nusbaum (MIT, Cambridge).

Best, C. T., McRoberts, G. W., and Sithole, N. N. (**1988**). "Examination of perceptual reorganization for nonnative speech contrasts: Zulu click perception by English-speaking adults and infants," J. Exp. Psychol. Hum. Percept. Perform. **14**, 345–360.

Best, C. T., and Strange, W. (**1992**). "Effects of phonological and phonetic factors on cross-language perception of approximants," J. Phonetics **20**, 305–330.

Bhat, D. N. S. (**1974**). "Retroflexion and retraction," J. Phonetics **2**, 233–237.

Bloch, B. (**1950**). "Studies in colloquial Japanese IV: Phonemics," Lang., **26**, 86–125.

Bradlow, A. R., (in press). "Training non-native language sound patterns: Lessons from training Japanese adults on the English /ɹ/-/l/ contrast," in *Phonology and Second Language Acquisition*, edited by M. Zampini and J. Hansen (Stanford University Press, Stanford, CA).

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., and Tohkura, Y. (**1999**) "Training Japanese listeners to identify English /r/ and /l/: Long term retention of learning in perception and production," Percept. Psychophys. **61**, 977–985.

Bradlow, A. R., Pisoni, D. B., Yamada, R. A., and Tohkura, Y. (**1997**). "Training Japanese listeners to identify English /r/ and /l/. IV. Some effects of perceptual learning on speech production," J. Acoust. Soc. Am. **101**, 2299–2310.

Carney, A. E., Widen, G. P., and Viemeister, N. F. (**1972**). "Noncategorical perception of stop consonants differing in VOT," J. Acoust. Soc. Am. **62**, 961–970.

Dave, R. (**1977**). "Retroflex and dental consonants in Gujarati: A palatographic and acoustic study," Annual Report of the Institute of Phonetics, University of Copenhagen Vol. **11**, pp. 27–156.

Delattre, P. (**1966**). *Comparing the Phonetic Features of English, French, German, and Spanish* (Harrap, London).

Dixit, R. P. (**1990**). "Linguotectal contact patterns in the dental and retroflex stops of Hindi," J. Phonetics **18**, 189–201.

Dixit, R. P., and Flege, J. E. (**1991**). "Vowel context, rate, and loudness effects on linguopalatal contact patterns in Hindi retroflex /t/," J. Phonetics **19**, 213–229.

Edman, T. R. (**1980**). "Learning of intraphonemic discrimination for several synthetic speech continua," Diss. Abstr. Int., B **41**, 4291B (University Microfilms No. 81-09,419).

Fairbanks, G. H., and Misra, B. G. (**1966**). *Spoken and Written Hindi* (Cornell University Press, New York).

Firth, J. R. (**1957**). "Word palatograms and articulation," in *Papers in Linguistics 1934–1951* (Oxford University Press, London).

Flege, J. E. (**1984**). "The effect of linguistic experience on Arabs' perception of the English /s/ vs. /z/ contrast," Folia Linguist. **18**, 117–138.

Flege, J. E. (**1989**). "Chinese subjects' perception of the word-final English /t/-/d/ contrast: Performance before and after training," J. Acoust. Soc. Am. **86**, 1684–1697.

Flege, J. E., and Eefting, W. (**1987**). "Production and perception of English stops by native Spanish speakers," J. Phonol. **15**, 67–83.

Flege, J. E., Takagi, N., and Mann, V. (**1996**). "Lexical familiarity and English-language experience affect Japanese adult's perception of /ɹ/ and /l/," J. Acoust. Soc. Am. **99**, 1161–1173.

Gillette, S. (**1980**). "Contextual variation in the perception of L and R by Japanese and Korean speakers," Minnesota Papers in Linguistics and the Philosophy of Language **6**, 59–72.

Goto, H. (**1971**). "Auditory perception by normal Japanese adults of the sounds 'L' and 'R'," Neuropsychologia **9**, 317–323.

Hardison, D. M. (**2003**). "Acquisition of second-language speech: Effects of visual cues, context, and talker variability," Appl. Psycholing. **24**, 495–522.

Hattori, S. (**1984**). *Onseigaku* [Phonology]. (Iwanami Shoten, Tokyo) (in Japanese).

Jamieson, D. G., and Morosan, D. E. (**1986**). "Training non-native speech contrasts in adults: Acquisition of the English /ϴ/-/ð/ contrast by francophones," Percept. Psychophys. **40**, 205–215.

Jamieson, D. G., and Morosan, D. E. (**1989**). "Training new non-native speech contrasts: A comparison of the prototype and perceptual fading techniques," Can. J. Psychol. **43**, 88–96.

Jenkins, J. J., and Yeni-Komshian, G. (**1995**). "Cross-language speech perception: Perspective and promise," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, edited by W. Strange (York, Timonium, MD).

Jones, D. (**1964**). *An Outline of English phonetics* (Heffer, Cambridge).

Jongman, A., Blumstein, S. E., and Lahiri, A. (**1985**). "Acoustic properties for dental and alveolar stop consonants: A cross-language study," J. Phonetics **13**, 235–251.
Jorden, E. H. (**1963**). *Beginning Japanese, Part 1* (Yale University Press, New Haven).
Ladefoged, P. (**1982**). *A Course in Phonetics* (Harcourt Brace Jovanovich, New York).
Ladefoged, P., and Bhaskararao, P. (**1983**). "Non-quantal aspects of consonant production: A study of retroflex consonants," J. Phonetics **11**, 291–302.
Levy E., and Strange, W. (**2002**). "Effects of consonantal context on perception of French rounded vowels by American English adults with and without French language experience," J. Acoust. Soc. Am. **111**, 2361.
Lisker, L., and Abramson, A. S. (**1970**). "The voicing dimension: Some experiments on comparative phonetics," in *Proceedings of the sixth International Congress of Phonetic Sciences* (Academia, Prague).
Lively, S. E., Logan, J. S., and Pisoni, D. B. (**1993**). "Training Japanese listeners to identify English /r/ and /l/. II. The role of phonetic environment and talker variability in learning new perceptual categories," J. Acoust. Soc. Am. **94**, 1242–1255.
Lively, S. E., Pisoni, D. B., Yamada, R. A., and Tohkura, Y. (**1994**). "Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories," J. Acoust. Soc. Am. **96**, 2076–2087.
Logan, J. S., Lively, S. E., and Pisoni, D. B. (**1991**). "Training Japanese listeners to identify English /r/ and /l/: A first report," J. Acoust. Soc. Am. **89**, 874–886.
Logan, J. S., Lively, S. E., and Pisoni, D. B. (**1993**). "Training listeners to perceive novel phonetic categories: How do we know what is learned," J. Acoust. Soc. Am. **94**, 1148–1151.
Logan, J. S., and Pruitt, J. S. (**1995**). "Methodological issues in training listeners to perceive non-native phonemes," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, edited by W. Strange (York, Timonium, MD).
MacKain, K. S., Best, C. T., and Strange, W. (**1981**). "Categorical perception of English /r/ and /l/ by Japanese bilinguals," Appl. Psycholing. **2**, 369–390.
Martin, S. E. (**1963**). *Essential Japanese* (Tuttle, Rutland, VT).
McCandless, B. D., Fiez, J. A., Protopapas, A., Conway, M., and McClelland, J. L. (**2002**). "Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception," Cog., Affect. & Behav. Neurosci. **2**, 89–108.
McClaskey, C. L., Pisoni, D. B., and Carrell, T. D. (**1983**). "Transfer of training of a new linguistic contrast in voicing," Percept. Psychophys. **34**, 323–330.
Miyawaki, K., Strange, W., Verbrugge, R. R., Liberman, A. M., Jenkins, J. J., and Fujimura, O. (**1975**). "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English," Percept. Psychophys. **18**, 331–340.
Moroson, D. E., and Jamieson, D. G. (**1989**). "Evaluation of a technique for training new speech contrasts: Generalization across voices, but not word-position or task," J. Speech Hear. Res. **32**, 501–511.
Ogawa, Y. (**1982**). *Nihongo Kyoiku Jiten* [Japanese Education Dictionary] (Taishuhkan Shoten, Tokyo) (in Japanese).
Pisoni, D. B., Aslin, R. N., Perey, A. J., and Hennessy, B. L. (**1982**). "Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants," J. Exp. Psychol. Hum. Percept. Perform. **8**, 297–314.
Polka, L. (**1989**). "The role of experience in speech perception: Evidence from cross-language studies with adults," Doctoral dissertation, University of South Florida (unpublished).
Polka, L. (**1991**). "Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions," J. Acoust. Soc. Am. **89**, 2961–2977.
Polka, L. (**1992**). "Characterizing the influence of native language experience on adult speech perception," Percept. Psychophys. **52**, 37–52.
Price, P. J. (**1981**). "A cross-linguistic study of flaps in Japanese and in American English," Doctoral dissertation, University of Pennsylvania (unpublished).
Pruitt, J. S. (**1992**). "Training native English speakers to identify Hindi retroflex-dental consonants," Masters thesis, University of South Florida (unpublished).
Pruitt, J. S., Strange, W., Polka, L., and Aguilar, M. (**1990**). "Effects of category knowledge and syllable truncation during auditory training on American's discrimination of Hindi retroflex-dental contrasts," J. Acoust. Soc. Am. **87**, S72.
Ramasubramanian, N., and Thosar, R. B. (**1971**). "Synthesis by rule of some retroflex speech sounds," Lang. Speech **14**, 65–85.
Rochet, B. L. (**1995**). "Perception and production of L2 speech sounds by adults," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, edited by W. Strange (York, Timonium, MD).
Shibatani, M. (**1990**). *The Languages of Japan* (Cambridge University Press, Cambridge).
Skinner, B. F. (**1954**). "The science of learning and the art of teaching," Harv. Educ. Rev. **24**, 86–97.
Skinner, B. F. (**1958**). "Teaching machines," Science **128**, 969–977.
Stevens, K. N., and Blumstein, S. E. (**1975**). "Quantal aspects of consonant production and perception: A study of retroflex stop consonants," J. Phonetics **3**, 215–233.
Strange, W. (**1972**). "The effects of training on the perception of synthetic speech sounds: Voice onset time," Doctoral dissertation, University of Minnesota (unpublished).
Strange, W. (**1992**). "Learning non-native phoneme contrasts: Interactions among subject, stimulus, and task variables," in *Speech Perception, Production, and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Eds.) (OHM, Tokyo, Japan), pp. 197–219.
Strange, W. (**1995**). "Cross-language studies of speech perception: An historical review," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, edited by W. Strange (York, Timonium, MD).
Strange, W., and Dittmann, S. (**1984**). "Effects of discrimination training on the perception of /r−l/ by Japanese adults learning English," Percept. Psychophys. **36**, 131–145.
Strange, W., Levy, E., and Lehnhoff, R. J., Jr. (**2004**). "Perceptual assimilation of French and German vowels by American English monolinguals: Acoustic similarity does not predict perceptual similarity," J. Acoust. Soc. Am. **115**, 2606.
Strange, W., Polka, L., and Aguilar, M. C. (**1989**). "Effects of auditory and phonetic training on American's discrimination of Hindi retroflex-dental contrasts," J. Acoust. Soc. Am. **86**, S101.
Svarny, O., and Zvelebil, K. (**1955**). "Some remarks on the articulation of the 'cerebral' consonants in Indian languages, especially in Tamil," Archiv Orientalni **23**, 374–434.
Tees, R. C., and Werker, J. F. (**1984**). "Perceptual flexibility: Maintenance or recovery of the ability to discriminate non-native speech sounds," Can. J. Psychol. **38**, 579–590.
Trehub, S. E. (**1976**). "The discrimination of foreign speech contrasts by adults and infants," Child Dev. **47**, 466–472.
Werker, J. F., and Tees, R. C. (**1984**). "Phonemic and phonetic factors in adult cross-language speech perception," J. Acoust. Soc. Am. **75**, 1866–1878.
Whalen, D. H., Wiley, E. R., Rubin, P. E., and Cooper, F. S. (**1990**). "The Haskins Laboratories pulse code modulation (PCM) system," Behav. Res. Methods Instrum. Comput. **22**, 550–559.
Yamada, R. A. (**1993**). "Effects of extended training on /r/ and /l/ identification by native speakers of Japanese," J. Acoust. Soc. Am. **93**, 2391.
Yamada, R. A. (**1995**). "Age effect on acquisition of non-native phonemes: Perception of English /r/ and /l/ for native speakers of Japanese," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, edited by W. Strange (York, Timonium, MD).
Yamada, R. A., Strange, W., Magnuson, J. S., Pruitt, J. S., and Clarke, W. D., III (**1994**). "The intelligibility of Japanese speakers productions of American English /r/, /l/, and /w/, as evaluated by native speakers of American English," Proceedings of the 1994 International Conference on Spoken Language Processing, pp. 2023–2026.

Perception of the [m]-[n] distinction in consonant-vowel (CV) and vowel-consonant (VC) syllables produced by child and adult talkers

Ralph N. Ohde
Department of Hearing and Speech Sciences, Vanderbilt Bill Wilkerson Center for Otolaryngology and Communication Sciences, Vanderbilt University Medical Center, 1215 21st Avenue South, Room 8310 Nashville, Tennessee 37232-8242

Katarina L. Haley
Division of Speech and Hearing Sciences, Department of Allied Health Sciences, Medical School Wing D, CB# 7190, University of North Carolina, Chapel Hill, North Carolina 27599-7190

Christine W. Barnes
Department of Hearing and Speech Sciences, Vanderbilt Bill Wilkerson Center for Otolaryngology and Communication Sciences, Vanderbilt University Medical Center, 1215 21st Avenue South, Room 8310 Nashville, Tennessee 37232-8242

(Received 3 January 2003; accepted 2 November 2005)

The contribution of the nasal murmur and vocalic formant transition to the perception of the [m]-[n] distinction by adult listeners was investigated for speakers of different ages in both consonant-vowel (CV) and vowel-consonant (VC) syllables. Three children in each of the speaker groups 3, 5, and 7 years old, and three adult females and three adult males produced CV and VC syllables consisting of either [m] or [n] and followed or preceded by [i æ u *a*], respectively. Two productions of each syllable were edited into seven murmur and transitions segments. Across speaker groups, a segment including the last 25 ms of the murmur and the first 25 ms of the vowel yielded higher perceptual identification of place of articulation than any other segment edited from the CV syllable. In contrast, the corresponding vowel+murmur segment in the VC syllable position improved nasal identification relative to other segment types for only the adult talkers. Overall, the CV syllable was perceptually more distinctive than the VC syllable, but this distinctiveness interacted with speaker group and stimulus duration. As predicted by previous studies and the current results of perceptual testing, acoustic analyses of adult syllable productions showed systematic differences between labial and alveolar places of articulation, but these differences were only marginally observed in the youngest children's speech. Also predicted by the current perceptual results, these acoustic properties differentiating place of articulation of nasal consonants were reliably different for CV syllables compared to VC syllables. A series of comparisons of perceptual data across speaker groups, segment types, and syllable shape provided strong support, in adult speakers, for the "discontinuity hypothesis" [K. N. Stevens, in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, edited by V. A. Fromkin (Academic, London, 1985), pp. 243–255], according to which spectral discontinuities at acoustic boundaries provide critical cues to the perception of place of articulation. In child speakers, the perceptual support for the "discontinuity hypothesis" was weaker and the results indicative of developmental changes in speech production. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2140830]

PACS number(s): 43.71.An, 43.71.Bp, 43.71.Es, 43.71.Ft [RLD] Pages: 1697–1711

I. INTRODUCTION

Several acoustic properties have been shown to contribute to the perception of place of articulation in adult and child speakers' production of nasal consonants (Kurowski and Blumstein, 1984; Repp, 1986; Repp and Svastikula, 1988; Ohde, 1994; Ohde and Perry, 1994; Ohde and Ochs, 1996). The most distinctive acoustic properties of consonant-vowel (CV) and vowel-consonant (VC) syllables include a low-frequency murmur representing sound passing through the nose and a relatively abrupt change in amplitude and spectrum at the juncture between the consonant and the vowel. Although listeners can identify place of articulation beyond chance from just the murmur (Kurowski and Blumstein, 1993), perceptual identification may be particularly accurate when based on tens of milliseconds of sound energy surrounding the point of spectral discontinuity between the consonant and vowel (Ohde, 1994; Ohde and Perry, 1994; Ohde and Ochs, 1996). In a study comparing the perception of nasals from murmur and vowel segments produced by three children and two adults, Ohde (1994) found that identification of place of articulation was more accurate from vowel segments than from murmur segments across the five speakers.

The perceptual importance of regions of rapid spectral and amplitude change has been discussed in the context of

quantal relations between speech physiology and acoustics (Stevens, 1989, 1998; Halle and Stevens, 1991). Quantal relations describe the common observation that there are relatively stable regions in which variations in articulation do not alter acoustic features, but that acoustic discontinuities can occur abruptly when the articulators change in degree of constriction from one stable region to another. The spectral change between nasal and vocalic resonance in the syllable [mi] is a good example of such acoustic discontinuities. Stevens (1985, 2002) suggested that corresponding abrupt changes in amplitude and spectrum play a critical role in perception by signaling boundary regions where important acoustic properties are located. As a result of several experiments, he concluded that the information in the boundary regions was robust and may provide invariant perceptual cues to place of articulation for oral and nasal stop consonants. Specifically, Stevens hypothesized that regions between 10 and 30 ms of boundaries separating consonant and vowel segments contain significant phonetic feature information, and that this information is accessed through perceptual auditory integration. Hereafter Stevens' conceptualization of the importance of spectral boundary cues to feature perception will be referred to as the "discontinuity hypothesis."

Investigators have suggested possible mechanisms and acoustic metrics underlying the perception of spectral discontinuities. Results from studies on nasal consonants indicate that the perceptual mechanisms involved in processing spectral change may include early integration of physiological processes mediated by short-term adaptation (Delgutte, 1980; Delgutte and Kiang, 1984) and/or late integration of cognitive processes mediated by long-term auditory memory (Repp, 1987; Ohde and Perry, 1994; Ohde and Ochs, 1996). The adaptation effect in humans may involve the murmur in CVs adapting the following vowel onset by attenuating low resonance frequencies and enhancing second and third formant transitions, where distinctive place of articulation cues are located. The acoustic metrics that have been proposed to account for the perception of spectral discontinuities in nasal consonant+vowel syllables include relational cues residing in the spectral change between the murmur and the vowel (Kurowski and Blumstein, 1987; Seitz *et al.*, 1990; Kurowski and Blumstein, 1993) and the relational property of the combination of separately processed murmur and vowel spectra (Harrington, 1994).

Perceptual research on cues to place of articulation identification for nasal consonants has been based, primarily, on studies employing CV syllable productions. Considerably less is known about variations in the perception of nasals in VC syllables. However, there is evidence indicating that the perceptual identification of consonants is more accurate (Repp and Svastikula, 1988; Redford and Diehl, 1999) and acoustic properties more robust (Ohala, 1990; Manuel, 1991; Wright, 2001) in CV syllables than in VC syllables.

Several factors make it difficult to predict the perceptual role of acoustic properties in nasal consonant VC syllables from research conducted on CV syllables. First, the transition between the vowel and murmur in VC syllables is not as abrupt as the murmur and vowel transition in CV syllables (Kurowski and Blumstein, 1987). In English, anticipatory velar lowering in VC syllables results in greater nasalization of the vowel than progressive velar lowering in CV syllables (Ali *et al.*, 1971). Second, data on locus equations indicate that stop consonant formant transitions are less distinctive in VC syllables than in CV syllables (Sussman *et al.*, 1997). Recent research employing locus equations has shown that formant transitions in nasal+vowel syllables are as distinct as those for stop+vowel productions (Hamby, 2000). Because nasals are produced similarly to their homorganic oral stops, it is reasonable to predict that formant transitions for nasals in VC syllables would be less distinctive than formant transitions in CV syllables (Sussman *et al.*, 1997). There is also perceptual research on stop consonants showing that listeners perform more poorly on place of articulation identification in the VC syllable position than in the CV syllable position (Wright, 2001). Third, differences in the acoustic correlates of place of articulation of nasals in VC compared to CV syllables may be greater in children than adults, because children do not articulate sounds as accurately in the VC syllable position as in the CV syllable position (Dyson, 1988; Stoel-Gammon, 1985; Ohde and Sharf, 1992). Thus, final consonant production may be more inexact or variable in children than in adults.

Although limited attention has been given to speaker factors, preliminary evidence suggests that place of articulation perception for nasal syllables is more difficult in some speakers than in others. Repp and colleagues (Repp, 1986; Repp and Svastikula, 1988) noted that identification of place of articulation was less accurate for nasals produced by a group of female talkers than for nasals produced by a group of male talkers. Similarly, Redford and Diehl (1999) observed clear differences among four adult speakers employed in a perceptual study of syllable position effects. Because studies on acoustic correlates of various sound features have demonstrated that interspeaker variability is greater for children than for adults (Ohde, 1985), identification of place of articulation may be less accurate for nasals produced by children than for nasals produced by adults. The reasons for these speaker variations should be examined (Klatt, 1986). According to Stevens' "discontinuity hypothesis," segments of the speech signal that contain regions of abrupt spectral change should be less sensitive to speaker variations than nonboundary regions with more static spectral properties.

The use of a perceptual level of analysis has been a very sensitive approach for the study of the development of speech production in children. There are clear anatomical and production differences between children and adults that may be examined perceptually (Parnell *et al.*, 1978; Ohde, 1994). For example, the relation of the larynx to the velum changes with age (Polgar and Weng, 1979; Kent and Murray, 1982), and the effect of this relationship may be greater nasalization in young children's speech than in older children's and adults' speech. If there is greater nasalization in children's speech compared to adults' speech, then it would be predicted that the spectral discontinuity cues at acoustic boundaries would be less important in children's speech than in adults' speech, because the potentially open velopharyngeal port would not permit the production of the temporally abrupt discontinuity.

The purpose of this research was to test the "discontinuity hypothesis" (Stevens, 1985). According to this hypothesis, information within a short time interval of the boundary (point of discontinuity) between the nasal and vowel contains the cues for accurate perception of place of articulation of this consonant. Implicit in the "discontinuity hypothesis" is the prediction that across speakers of a specific age and gender, place of articulation of nasals will be more accurately identified from speech segments including these boundary regions than from any other speech segments. The strong predictions of the hypothesis are that short duration boundary segments should provide the essential acoustic information used to perceive consonant place of articulation and that the relative importance of this segment should hold across variations in speaker group, speaker proficiency, and syllable type. The following questions were addressed: (1) Do spectral discontinuities at consonant-vowel boundaries provide the essential information to perceive nasal place of articulation accurately across speaker group, talkers within a speaker group, and syllable position as predicted by the "discontinuity hypothesis"? (2) Is the CV syllable perceptually distinctive compared to the VC syllable? (3) Are there talker and developmental differences for the murmur and transition properties as evidenced in the perception and production of these cues to place of articulation?

II. EXPERIMENT I: THE PERCEPTUAL DISTINCTION OF [m]-[n] IN CV SYLLABLES

A. Method

1. Talkers and recording procedures

The speech sample consisted of nasal consonant + vowel syllables produced by three male adults, three female adults, and three children from each of the speaker groups 3, 5, and 7. All talkers were judged to speak a General American dialect on the basis of informal observation. The children were between −1 and +5 months of their designated age. They were required to pass a hearing screening at 20 dB for the octave frequencies between 250 and 8000 Hz. In addition, they performed within normal limits on either the *Fluharty Preschool Speech and Language Screening Test* (Fluharty, 1974, 1978) or the *Bankson Language Screening Test* (Bankson, 1977), and on the *Arizona Articulation Proficiency Scale-Revised* (Fudala, 1970). All recordings were performed in a sound-treated IAC booth with a high-quality recorder (Revox, Model B 77 MKII) and microphone (Neumann, Model TLMI70i), placed approximately 10 in. from the talker's mouth. The frequency response of the microphone was flat out to about 20 kHz. Five repetitions of the syllable combinations of [m n] and [i æ u *a*] were produced in a fixed random order. The adults read the syllables from a list and the children repeated the syllables after one of the experimenters. Although some of these syllables were meaningful (e.g., [ni]) and others were nonsense (e.g., [næ]), neither adults nor children experienced any difficulty producing these syllable types. As reported in the results section, the mean percent identification of these syllables was very high for child and adult talkers supporting accurate and appropriate articulation. The productions were digitized at a sampling rate of 22 kHz and stored on a PDP 11/73 computer. A 12-bit quantization was used, and the output was low-pass filtered at 10 kHz.

The 15 talkers were selected from a larger set of talkers in a manner that intended to reflect a range of clarity of production for the test syllables, as would be expected in a typical population. The selection was based on a preliminary pilot test, in which five adults rated nasal consonant + vowel syllables produced by six talkers in each speaker group in terms of the clarity of the perceived place of articulation (Ohde, 1994). In the current investigation, talkers refer to the individual participants in each speaker group. Thus, in this study we are sometimes interested in differences among individual talkers for a speaker group, e.g., 3-year-old group, and at other times we are interested in speaker group differences collapsed across talkers within these groups. All participants were native speakers of English and reported no history of speech or hearing disorder. They were paid on an hourly basis for their participation. A four-point rating scale was used, where a level of 1 represented "good m," 2 "not so good m," 3 "not so good n," and 4 "good n." The mean rating was calculated across five productions of each syllable. Ambiguous productions were operationally defined as an average rating between 1.8 and 2.2 for [m] and between 2.8 and 3.2 for [n]. Within each age group, one talker with a small number, one with an intermediate number, and one with a relatively high number of ambiguous utterances were selected. As will be reported in the Results section, the full syllable was still identified at or above 95% for all talkers, which indicates that place of articulation was well preserved for all talkers. None of the selected talkers was used in the previous study (Ohde, 1994).

2. Stimulus preparation

The first and the fifth production of each syllable were used for preparation of the experimental stimuli. Eight stimulus segments were generated through the use of a waveform-editing program. For each syllable, the point of discontinuity between the murmur and the vowel was visually located on the waveform. If there was a question about the location of this point, supplementary broadband spectrographic analysis was performed and a decision was made on the basis of the combined information from visual inspection of the waveform, spectrographic analysis, and auditory perception.

In Fig. 1, a spectrographic representation of all eight segments is provided for an adult male talker. After the point of discontinuity (Fig. 1—segmentation mark [SM] 4) had been determined, each syllable production was edited into the following eight segment types: (1) full murmur (FM)—the period of oral closure from syllable onset up to and including the last low amplitude murmur period preceding the discontinuity (SMs 1–4); (2) 50-ms murmur (50M)—the last 50 ms of the murmur (SMs 2–4); (3) 25-ms murmur (25M)—the last 25 ms of the murmur (SMs 3–4); (4) 25-ms murmur + 25-ms transition (MT)—the last 25 ms of the murmur and the first 25 ms of the immediately following vowel (SMs 3–5); (5) 25-ms transition (25T)—the first 25 ms of the vowel (SMs 4–5); (6) 50-ms transition—the first 50 ms of the vowel (SMs 4–6); (7) full transition

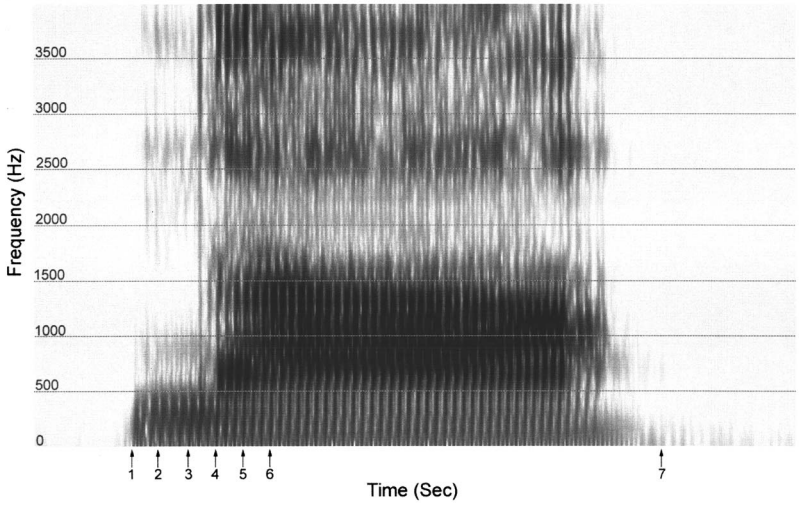


FIG. 1. A spectrographic illustration of an adult male's production of [n*a*]. The segmentation marks (SMs) represent the following eight stimulus conditions of experiment I: full murmur (FM-SMs 1 to 4), 50-ms murmur (50M-SMs 2 to 4), 25-ms murmur (25M-SMs 3 to 4), 25-ms murmur+25-ms transition (MT-SMs 3 to 5), 25-ms transition (25T-SMs 4 to 5), 50-ms transition (50T-SMs 4 to 6), full transition (TV-SMs 4 to 7), and full syllable (FS-SMs 1 to 7).

(TV)—from the discontinuity to the offset of the syllable including the formant transition and the quasi-steady-state of the vowel (SMs 4–7); and (8) full syllable (FS)—the complete syllable including the full murmur and the full transition (SMs 1–7). All cuts were made at zero-crossings. Cuts within the murmur were made on the downslope of a period, and cuts within the vowel were made on the upslope. An attempt was made to keep the edited segments as close to their nominal values of 25 and 50 ms as possible. Because there was intertalker variability in f_0 and the position of the cut within a period was consistent, the actual duration of the segments varied somewhat. However, this variation never exceeded ±2 ms for the child productions and ±5 ms for the adult productions. In editing the MT segments, an additional criterion was that an equal number of periods were included from the vowel and from the murmur.

3. Participants and procedures

The participants were ten graduate students in the Hearing and Speech Sciences program at Vanderbilt University. All were native speakers of English and reported no history of speech or hearing disorder. They were paid on an hourly basis for their participation.

The testing was conducted in an IAC booth over matched and calibrated headphones (TDH-49) at a comfortable listening level of approximately 80 dB SPL. The stimulus presentation and data collection were controlled online by a PDP 11/73 computer. The listeners were seated in front of a response box with two buttons labeled "m" and "n" and responded by pressing one of these buttons in a two-alternative forced-choice (2AFC) testing paradigm. They were instructed to respond to each stimulus by pressing the button for the nasal consonant with the corresponding place of articulation and to guess if they were not sure. The interstimulus interval was constant at 3 s. Typically 128 stimuli (2 consonants × 4 vowels × 2 productions × 8 stimulus conditions) were generated for each of the 15 speakers. For some speakers, there were slightly fewer stimuli, because some productions did not contain a sufficiently long murmur for the 50-ms murmur edit (Fig. 1—SMs 2–4). Each stimulus was presented twice.

Some stimulus variables were grouped within blocks and others were randomized. The five speaker groups were presented on separate sessions. The three talkers within each of these groups were presented in separate blocks during a session, and the eight stimulus segments were blocked within each talker. The 32 stimuli within each of the segment type blocks (2 consonants × 4 vowels × 2 productions × 2 repetitions) were presented in random order. The different segment types were presented in succession, starting with the full syllable followed by the edited segments. The order of the murmur and vowel segments alternated between tests, but the longer stimuli for a given segment type were always presented before the shorter stimuli. The order of presentation for speaker groups, individual talkers within the groups, and murmur and vowel segments was counterbalanced across listeners.

4. Statistical analyses

Repeated measures analyses of variance on the arcsine transformed percents were used to analyze the data. In these analyses, overall effects of speaker group and talker were computed as in previous research (Repp, 1987; Ohde, 1994; Ohde and Perry, 1994; Ohde and Ochs, 1996). Tests of simple main effects and appropriate Newman Keuls' *post hoc* tests were used to examine significant interactions. The results of simple main effects analyses are reported in footnotes. Significant results at the $p < 0.05$ or $p < 0.01$ levels are reported. Some marginally significant findings at either $p < 0.06$ or $p < 0.10$ are also reported.

B. Results and discussion

1. Effects of speaker group

The mean whole syllable identification ranged from 95% for speech produced by 3- and 5-year olds to 99% for speech produced by adult females and males. Whole syllable identification for speech produced by 7-year olds was 97%.

a. Murmur cue. As illustrated in Fig. 2 (panel a—FAD: female adults; MAD: male adults), the murmur was identified with correct place of articulation at an above-chance level across speaker groups. However, it is also very appar-

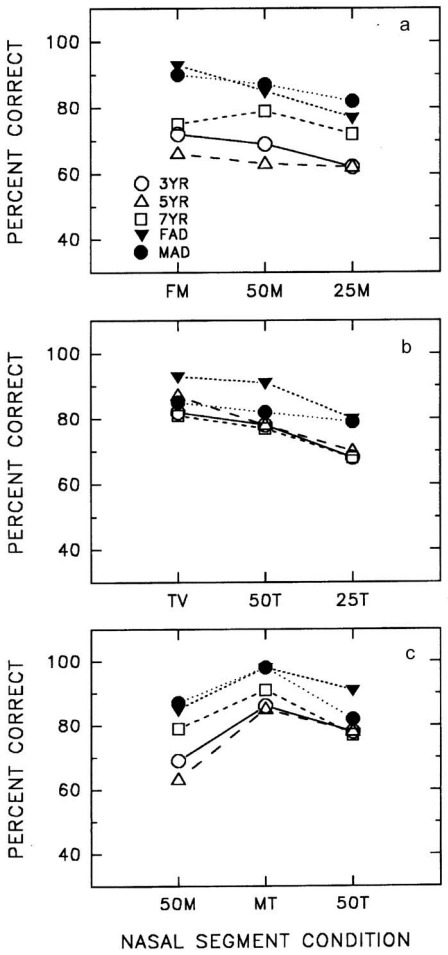


FIG. 2. Mean percent correct identification of nasal consonants from the full murmur (FM), 50-ms murmur (50M), and 25-ms murmur (25M) stimuli (panel a), from the full transition and vowel (TV), 50-ms transition (50T), and 25-ms (25T) transition stimuli (panel b), and from the 50-ms murmur (50M), 25-ms murmur+25-ms transition (MT), and 50-ms transition (50T) stimuli (panel c) collapsed across talkers within a speaker group (FAD: female adults; MAD: male adults).

ent that substantial differences exist across speaker groups. The identification of place of articulation was more accurate for speech produced by adult speakers than for speech produced by 3- and 5-year-old speakers. An analysis of variance revealed significant effects for speaker group $[F(4,36)=63.82;\ p<0.01]$, stimulus duration $[F(2,18)=34.16;\ p<0.01]$, and speaker group×stimulus duration interaction $[F(8,72)=5.06;\ p<0.01]$. *Posthoc* tests showed that with the exception of the 5-year-old group, a significant duration effect was found in each speaker group.[1] There was a direct relation between the duration of the murmur and the accuracy of identification of the nasal. For example, as shown in Fig. 2, identification of place of articulation of the murmur was more accurate for the 50M segment than the 25M segment. This trend in perception was somewhat different for the 7-year-old speaker group in that the accuracy in the identification of the nasal from the FM segment was poorer than from the 50M segment. The speaker group effect was significant across murmur duration.[2] Identification of nasals from adult productions was significantly more accurate than from all 3- and 5-year-old children's productions, and perception of nasals from the 7-year-old group's productions was significantly more accurate than those from the 3-year- and 5 -year-old speaker groups' productions.

b. Transition cue. The results for the coarticulated transition stimuli are shown in Fig. 2 (panel b). Unlike the findings for the murmur, identification of the nasals from transitions was not consistently more accurate for adult speakers than for child speakers. However, for the shortest vowel durations, there was a significant advantage for adult speakers. An analysis of variance showed that speaker group $[F(4,36)=18.94;\ p<0.01]$, stimulus duration $[F(2,18)=120.61;\ p<0.01]$, and the speaker group×stimulus duration interaction $[F(8,72)=3.73;\ p<0.01]$ were significant. *Posthoc* tests for the transition stimuli established that duration was significant for all speaker groups.[3] Specifically, as the transition duration increased there was a consistent increase in the accuracy of place of articulation identification of nasals. As was the case for the murmur, developmental effects of speaker group were observed for the 25T stimulus condition. For this condition, identification of nasals was significantly more accurate for adult speaker groups than child speaker groups. Thus, the strength of the transition property was linked to stimulus duration, providing a stronger cue in adult speech than in child speech at short stimulus durations. For the longer 50T and TV segments, identification of nasals was significantly more accurate for the female adult speaker group than for the child speaker groups. Also, at the longer transition segment durations, identification of nasals was more accurate from the adult female speaker group than from the adult male speaker group, which is contrary to the predictions of earlier research[4] (Repp, 1986; Repp and Svatiskula, 1988). Finally, a comparison of panels (a) and (b) for the child groups shows that identification was more accurate from transition stimuli than from murmur stimuli, a finding not observed for adult groups. It appears that the acoustic correlates of the murmur for place of articulation are not well represented in child-speaker groups compared to adult-speaker groups. The relative strength of the transition cue in the child-speaker groups may relate to the F2 locus that is well represented in children's productions according to recent developmental research on locus equations in nasals (Hamby, 2000).

c. Murmur+transition versus murmur and whole syllable cues. To examine the perceptual identification of the murmur+transition stimulus relative to speaker group and segment type, an analysis of variance was conducted on the MT segment in comparison to the murmur segments and to the whole syllable. If the MT segment illustrated in Fig. 2 (panel c) is the dominant cue in perception of prevocalic nasals as predicted by the "discontinuity hypothesis," then there should be a significant segment type effect. Specifically, identification of nasal place of articulation should be significantly more accurate from the MT segment than from all the murmur segments (panels a and c) and if the spectral discontinuity contains all the essential information for place of articulation perception, there should be no difference in the accuracy of identification of nasals from the MT segment compared to the whole syllable. The results for the MT and murmur segments and whole syllable showed significant ef-

fects for speaker group [$F(4,36)=86.61$; $p<0.01$], segment type [$F(4,36)=133.65$; $p<0.01$], and speaker group $\times$segment type interaction [$F(16,144)=11.23$; $p<0.01$]. According to *posthoc* tests, the perceptual effect of segment type was significantly different across all speaker groups,[5] and identification of nasal place of articulation was significantly more accurate from the MT segment than from any of the murmur segments (25M, 50M, FM) for all speaker groups. The prediction that identification would be similar for the MT segment and the whole syllable was met for the adult speaker groups, but not for the child speaker groups. For each of the child speaker groups, identification of nasals was significantly higher from the whole syllable than from the MT segment. Thus, it appears that the perceptual salience of the MT segment evolves with age and is not complete by even 7 years of age.

d. Murmur+transition versus transition and whole syllable cues. The results for the MT versus transition segments and whole syllable productions revealed significant speaker group [$F(4,36)=35.32$; $p<0.01$] and segment type [$F(4,36)=180.27$; $p<0.01$] main effects, and a significant speaker group$\times$segment type interaction [$F(16,144)=6.31$; $p<0.01$]. As predicted by the "discontinuity hypothesis," *posthoc* tests revealed that nasal identification from the MT segment was significantly more accurate than from any of the transition segments (25T, 50T, TV) [Fig. 2 (panels b and c)]. This was true for all speaker groups except the 5-year-old group,[6] where identification was not significantly higher from the MT segment than from the TV segment.

In summary, the results for our adult speakers provide strong support for the "discontinuity hypothesis." Identification of nasals from the adults' productions of MT segments across vowels and consonants was near 100%, indicating that additional acoustic information provided in the vowel and murmur segments was perceptually redundant to the identification of nasal place of articulation. Although place of articulation was also perceived with a high degree of accuracy in the MT segments of children's syllable production, the dominance of this syllable segment was less clear, suggesting a developmental process.

2. Within group talker effects

Recall that the talkers selected in this research represented a range of quality of production at each age level. Thus, we would predict that the accuracy in identification of nasals from the talker with the best rating (talker 1) for place of articulation would be higher than from the talker with the poorest rating (talker 3). To examine specific talker effects, analyses were performed on the three talkers (T1, T2, and T3) for the 50-ms segment types (50M, MT, 50T). The prediction based on the "discontinuity hypothesis" would be significantly more accurate identification of the nasal from the MT segment than from either the 50M or 50T segments across talkers.

The results of these analyses for the three stimulus segments revealed significant main effects of segment type in all speaker groups (3-year: [$F(2,18)=42.28$; $p<0.01$]; 5-year: [$F(2,18)=41.46$; $p<0.01$]; 7-year: [$F(2,18)=22.71$; $p<0.01$]; FAD—female adult: [$F(2,18)=14.22$; $p<0.01$]; MAD—male adult: [$F(2,18)=20.81$; $p<0.01$]). There was a main effect for speaker rating in all groups, except the 3-year old (5-year: [$F(2,18)=15.30$; $p<0.01$]; 7-year: [$F(2,18)=3.56$; $p<0.05$]; FAD: [$F(2,18)=7.46$; $p<0.01$]; MAD: [$F(2,18)=17.46$; $p<0.01$]). The speaker rating $\times$segment type interaction was significant in all groups except the 7-year old ([3-year: $F(4,36)=5.67$; $p<0.01$]; [5-year: $F(4,36)<4.96$; $p<0.01$]; [FAD: $F(4,36)=5.20$; $p<0.01$]; [MAD: $F(4,36)=11.28$; $p<0.01$]). *Posthoc* tests revealed that for the 5-year-old, 7-year-old, FAD, and MAD speaker groups, identification of place of articulation was significantly more accurate from the MT segment than from either the 50M or 50T segments for 7 of the 12 talkers (5-year old: T2; 7-year old: T1, T2, T3; FAD: T3; MAD: T2, T3), and marginally significant ($p<0.10$) for two other talkers (5-year old: T3; FAD: T2).[7] It is also important to note that identification of the nasal from the MT segment was significantly more accurate than from either the 50M or 50T segments for talker 3, who had the poorest pretest quality rating for all these older groups. It may be that listeners actively detect multiple cues for a particular feature, and when specific cues are deficient an integrated cue is critical for feature perception. In the current case, the combination of the murmur+transition forming the prominent spectral discontinuity was clearly important for the perceptual identification in talkers. For the young 3-year-old speaker group, more accurate identification of the MT segment than the 50M or 50T segments occurred for one of the three talkers. Interestingly, this talker was also talker 3. These findings indicate that the motor control required to produce spectral change is learned over sound development and represents a process that is still emerging in 3-year-old children.

Across talkers, identification of the nasals was also less variable from the MT segment than from either the murmur or transition segments. The standard deviation in arcsine percents averaged across talkers for the MT, 50M, and 50T segments was 0.0770, 0.1484, and 0.1232, respectively. Thus, perceptual variability for the MT segment was substantially less than the variability observed for the 50M and 50T segments. An ANOVA and Newman Keuls' *posthoc* tests revealed that the MT segment was significantly [$F(2,14)=14.31$; $p<0.01$] less variable than the 50T and 50M segments, and that the 50M and 50T segments were not significantly different. As illustrated in Fig. 3 (panels a and b), within a particular speaker group, identification of the nasal from the murmur and transition could vary by over 20% across talkers. For example, T1 compared to T3 for the murmur (panel a) in the MAD speaker group, and T1 compared to T2 for the transition (panel b) in the 5-year-old speaker group. On the other hand, nasal identification was very consistent across talkers for the MT segment [Fig. 3 (panel c)].

To summarize, experiment I allowed us to test, in CV syllables, the predictions of the "discontinuity hypothesis" according to variations in speaker group, talker, and segment condition. According to this hypothesis, identification of nasals in CV syllables will be more accurate from the MT segment than from the murmur and transition segments across speaker groups and talkers. In general, the results support the hypothesis. With the exception of the 3-year olds,

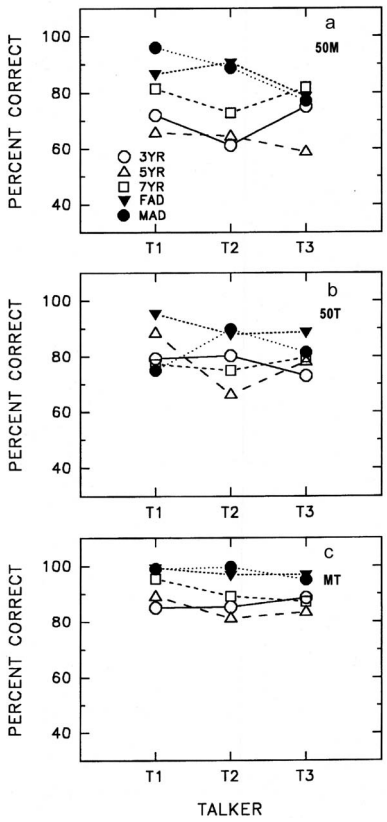


FIG. 3. Mean percent correct identification of nasal consonants from the 50-ms murmur (50M) (panel a), the 50-ms transition (50T) (panel b), and the 25-ms murmur+25-ms transition (MT) stimuli (panel c) as a function of talker within a speaker group (FAD: female adults; MAD: male adults).

our perceptual results indicate that the region of spectral discontinuity at the junction between the murmur and the vowel contained more important cues to nasal place of articulation than any other segment of the consonant-vowel sequence. For the 3-year-old speaker group, motor processes involved in producing spectral discontinuities appear to be still emerging in CV syllables. The analysis of talkers within speaker groups provided further support for the spectral discontinuity in the MT stimulus as the dominant cue for nasal identification and for the development of its perceptual importance in children.

III. EXPERIMENT II: THE PERCEPTUAL DISTINCTION OF [m]-[n] IN VC SYLLABLES

A. Method

1. Stimuli

The same talkers were used as in experiment I. It was deemed critical to use these talkers as a direct internal control in speech production for the comparison of the perceptual results of these talkers' CV and VC syllables. The VC productions of one 3-year-old talker did not contain murmur segments. This talker, who was rated second in terms of clarity of articulation of nasal place of articulation in CV syllables (see experiment I), was therefore not included in the experiment. The recording and digitizing procedures were the same as in experiment I. The talkers produced VC syllables consisting of the vowels [i æ u *a*] followed by the nasal consonants [m n]. As in experiment I, five productions of each syllable were recorded and the first and fifth productions were selected for the experiment. The computer editing of the waveform was analogous to the editing of the CV syllables. Figure 4 shows a spectrographic illustration of all eight segments for an adult male talker. Numbers go from right to left so that for a given segment they are the same numerically in the VC and CV syllables. After locating the point of discontinuity (Fig. 4—SM 4) between the vowel and the murmur, the following eight segment types were generated from each syllable: (1) full transition (TV)—from syllable onset to the discontinuity between the vowel and the murmur including the quasi-steady-state of the vowel and the formant transition (SMs 4–7); (2) 50-ms transition—the last 50 ms of the vowel (SMs 4–6); (3) 25-ms transition (25T)—the last 25 ms of the vowel (SMs 4–5); (4) 25-ms transition+25-ms murmur (MT)—the last 25 ms of the

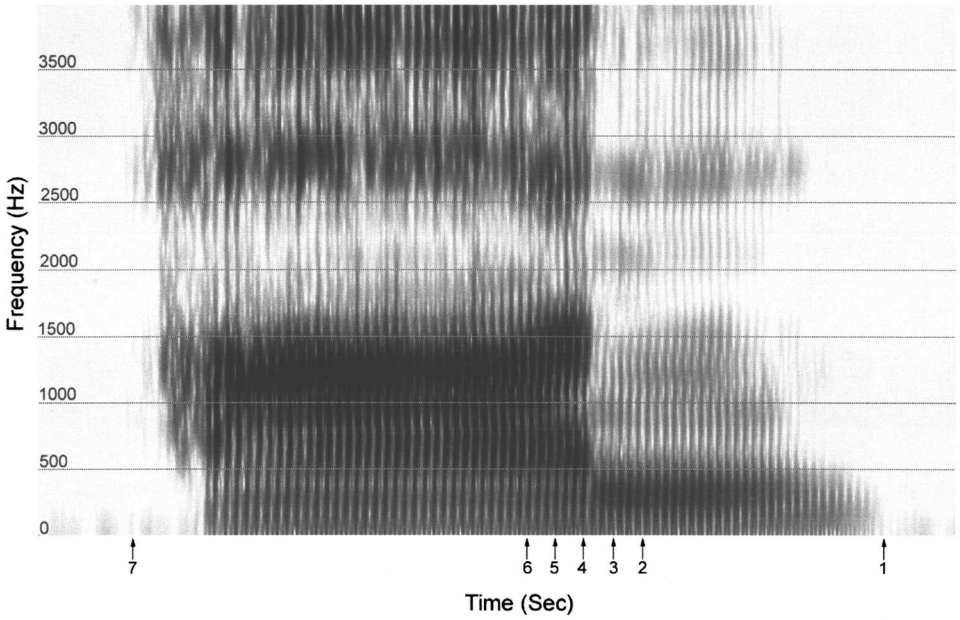


FIG. 4. A spectrographic illustration of an adult male's production of [*a*n]. The segmentation marks (SMs) represent the following eight stimulus conditions of experiment II: full murmur (FM-SMs 1 to 4), 50-ms murmur (50M-SMs 2 to 4), 25-ms murmur (25M-SMs 3 to 4), 25-ms murmur +25-ms transition (MT-SMs 3 to 5), 25-ms transition (25T-SMs 4 to 5), 50-ms transition (50T-SMs 4 to 6), full transition (TV-SMs 4 to 7), and full syllable (FS-SMs 1 to 7).

vowel and the first 25 ms of the murmur (SMs 3–5); (5) 25 -ms murmur (25M)—the first 25 ms of the murmur (SMs 3–4); (6) 50-ms murmur (50M)—the first 50 ms of the murmur (SMs 2–4); (7) full murmur (FM)—from the discontinuity at nasal closure to the end of the syllable (SMs 1–4); and (8) full syllable (FS)—the complete syllable including the full transition and the full murmur (SMs 1–7). All cuts were made at zero crossings. Cuts were made on the downslope of periods within the vowel and on the upslope within the murmur. The variations from the nominal durations of 25 and 50 ms never exceeded ±4 ms for the child productions and ±6.5 ms for the adult productions. As was the case for the CV syllables, the MT segments were symmetrical in terms of the number of vowel and murmur periods.

2. Participants and procedures

The participants were ten adults who had the same general characteristics as those in experiment I, but who did not participate in that study. The presentation procedures and ordering principles were the same as in experiment I, with the one exception that stimuli from only two talkers in the 3-year-old group were used. For each of the 14 talkers, VC murmurs were greater than 50 ms and accommodated all edits.

3. Statistical analyses

As in experiment I, repeated measures analyses of variance on the arcsine transformed percents were used to analyze the data. The statistical analysis procedures were identical to experiment I.

B. Results and discussion

1. Effects of speaker group

The mean whole syllable identification ranged from 86% for speech produced by 3-year olds to 98% for speech produced by adult males. Whole syllable identification was 93% for speech produced by 7-year olds and 94% for speech produced by 5-year olds and female adults.

a. Murmur cue. As shown in Fig. 5 (panel a) (FAD: female adults; MAD: male adults), identification of place of articulation of the murmur was highest for segments produced by male adults followed by female adults and then children. An analysis of variance showed that speaker group $[F(4,36)=18.78;\ p<0.01]$ and stimulus duration $[F(2,18)=5.46;\ p<0.05]$ were significant. Newman Keuls' *posthoc* tests confirmed that place of articulation identification from the nasal murmur was significantly more accurate from the adult male speaker group than from any of the other speaker groups including the female adults. Although listeners were relatively poor in their identification of nasals from the murmur produced by female adults, the identification of nasals was significantly more accurate from the FAD speaker group than from the 3- and 5-year-old speaker groups. These findings for the VC syllable are very different from those for the CV syllable where nasal identification was high and comparable for the two adult speaker groups (MAD and FAD). They still provide support for developmental differences between adults and young children. In regard to the main effect of duration, Newman Keuls' *posthoc* tests revealed that identification of nasals was significantly more accurate from the FM and 50M segments than from the 25M segment.

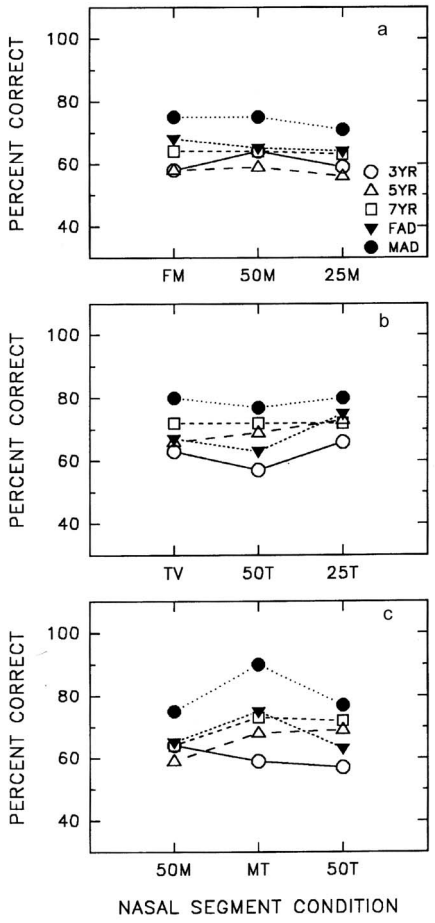


FIG. 5. Mean percent correct identification of nasal consonants from the full murmur (FM), 50-ms murmur (50M), and 25-ms murmur (25M) stimuli (panel a), from the full transition and vowel (TV), 50-ms transition (50T), and 25-ms (25T) transition stimuli (panel b), and from the 50-ms murmur (50M), 25-ms murmur+25-ms transition (MT), and 50-ms transition (50T) stimuli (panel c) collapsed across talkers within a speaker group (FAD: female adults; MAD: male adults).

b. Transition cue. As illustrated in Fig. 5 (panel b), the transition was perceived above chance level across speaker groups. However, it was also very apparent that substantial differences existed for speaker groups. For the transition stimulus conditions, the identification of place of articulation of the nasal was generally highest from the adult male speaker group and lowest from the 3-year-old and female adult speaker groups. An analysis of variance revealed significant effects for speaker group $[F(4,36)=28.54;\ p<0.01]$, stimulus duration $[F(2,18)=11.04;\ p<0.01]$, and the speaker group×stimulus duration interaction $[F(8,72)=3.51;\ p<0.01]$. Contrary to the findings for the CV transition stimuli, *posthoc* tests for the VC transition duration conditions revealed significant segment duration effects for only the 3-year-old and FAD groups.[8] Interestingly, duration was inversely related to the accuracy of nasal identification for these groups, i.e., as duration decreased, correct perception of place of articulation of nasals increased. It is possible that this duration finding is due to a form of forward masking of

the vowel steady state on the transition. As Fig. 4 shows, the 25T segment contains the important transition cue minus any preceding vowel, whereas both the 50T and the TV segments contain vowel steady-state energy that could mask the vowel formant transition. On the other hand, it is unclear why masking would influence perception of productions from only two of the five speaker groups. In addition, across duration, the only consistent effect was that the identification of stimuli generated from adult male (MAD) utterances was significantly more accurate than those generated from all other speaker groups. Thus, developmental effects were not as apparent for the VC transition stimuli as they were for the comparable CV conditions.

c. Murmur+transition versus murmur and whole syllable cues. Two analyses of variance were conducted on the MT segment in comparison to the murmur and transition segments. Similar to the findings for the CV syllable conditions, the results for the MT and murmur segments [Fig. 5 (panels a and c)] revealed significant effects for speaker group [$F(4,36)=32.53$; $p<0.01$], segment type [$F(4,36)=368.17$; $p<0.01$], and speaker group × segment type interaction [$F(16,144)=7.05$; $p<0.01$]. *Posthoc* tests showed that the perceptual effect of segment type was significantly different across all speaker groups, and that identification of nasal place of articulation was significantly more accurate from the whole syllable than from the MT segment or any of the murmur segments (25M, 50M, and FM).[9] With the exception of the 3-year-old speaker group, identification of nasal place of articulation was significantly more accurate from the MT segment than from any of the murmur segments.

d. Murmur+transition versus transition and whole syllable cues. Contrary to the findings for the CV syllable transition conditions, the results for the VC syllable MT and transition segments were variable [Fig. 5 (panels b and c)]. There were significant speaker group [$F(4,36)=49.69$; $p<0.01$] and segment type [$F(4,36)=153.58$; $p<0.01$] main effects, and a significant speaker group × segment type interaction [$F(16,144)=6.31$; $p<0.01$]. However, *posthoc* tests showed that nasal identification from the MT segment was not significantly more accurate than identification from the transition segments (25T, 50T, TV) for the 3-, 5-, and 7-year-old groups. As illustrated in Fig. 5, nasal identification was significantly more accurate from the MT segment than from all the transition segments for only the FAD and MAD groups, with the exception of the FAD MT versus 25T segment condition.[10] Across all speaker groups, identification of nasal place of articulation was significantly more accurate from the whole syllable than from the MT segment and from any of the transition segments (25T, 50T, TV).

In summary, these analyses of speaker groups show that the MT stimulus was not the dominant cue for identification of place of articulation of nasals in the VC context. For all speaker groups, identification of nasals was significantly more accurate from the whole syllable than from the MT segment and from any of the transition segments (25T, 50T, TV). For 11 of the 12 adult speaker group comparisons, the accuracy of place of articulation identification of the nasal was significantly greater from the MT segment than from the murmur and transition segments, but no such differences were found in the child speaker groups. As such, these findings for VC syllable production support the "discontinuity hypothesis" for adult speakers and suggest that important qualitative changes occur in children's vowel+nasal production up to at least 7 years of age.

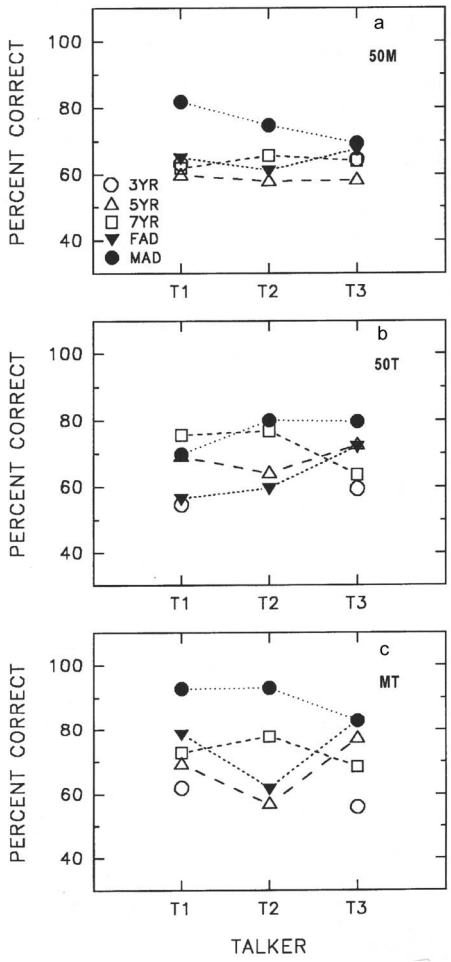


FIG. 6. Mean percent correct identification of nasal consonants from the 50-ms murmur (50M) (panel a), the 50-ms transition (50T) (panel b), and the 25-ms murmur+25-ms transition (MT) stimuli (panel c) as a function of talker within a speaker group (FAD: female adults; MAD: male adults).

2. Within group talker effects

To examine specific talker effects, analyses similar to those for CV syllables were performed on the three talkers (T1, T2, T3) for the 50-ms segment types (50M, MT, 50T) that are illustrated in Fig. 6 (only talkers T1 and T3 for the 3-year-old group). The prediction based on the "discontinuity hypothesis" would be significantly more accurate identification of the nasal from the MT segment than from either the 50M or 50T segments across talkers. However, as indicated in the Introduction, it is quite possible that the MT segment in the VC syllable does not carry the same perceptual weight as it does in the CV syllable.

The results of these analyses for the three stimulus conditions revealed significant main effects of segment type (5-year: [$F(2,18)=10.81$; $p<0.01$]; 7-year: [$F(2,18)=8.20$; $p<0.01$]; FAD: [$F(2,18)=28.99$; $p<0.01$]; MAD: [$F(2,18)=26.61$; $p<0.01$]) and talker (5-year: [$F(2,18)$

$=10.98$; $p<0.01$]; 7-year: [$F(2,18)=11.54$; $p<0.01$]; FAD: [$F(2,18)=17.16$; $p<0.01$]; MAD: [$F(2,18)=4.69$; $p<0.05$]) for all speaker groups, except for the 3-year old. The talker × segment interaction was significant for the 5-year-old and the adult speaker groups ([5-year: $F(4,36)=3.34$; $p<0.05$]; [FAD: $F(4,36)=6.69$; $p<0.01$]; [MAD: $F(4,36)=5.84$; $p<0.01$]). *Posthoc* tests revealed that identification of place of articulation was significantly more accurate from the MT segment than from either the 50M or 50T segments for only four talkers.[11] Two were from the FAD (T1 and T3) group and two were from the MAD group (T1 and T2). These findings provide further developmental support for the emergence of the spectral discontinuity property as a function of talker age. Unlike the findings for the CV syllable, nasal identification from the VC syllable MT segment was significantly more accurate than nasal identification from either the 50M or 50T segments for only FAD talker 3. Recall that talker 3 had the poorest pretest quality rating for CV productions. For the CV syllable, the MT segment was identified more accurately than the 50M and 50T segments for talker 3 from all speaker groups. These findings indicate that integrated cues for feature perception for talkers representing a range of perceptual quality may be linked to syllable context. The CV syllable appears to represent the perceptually dominant context from which features can be perceived from either individual cues such as the transition or integrated cues such as the murmur+transition. For speakers who produce sounds with relatively poor quality, integrated cues seem to be particularly important for feature identification and these cues are more readily available from CV syllables than from VC syllables.

Another important difference between the VC and CV syllable was in the standard deviation of the arcsine percents averaged across talkers for the perception of the MT, 50M, and 50T segments (0.1249, 0.1404, and 0.1432, respectively). Although talker variability was smallest for the MT segment, the differences among stimulus types was not statistically significant. A comparison of Figs. 3 (panel c) and 6 (panel c) illustrates these differences in variability of nasal perception across talkers for the MT segment in CV and VC syllables.

To summarize, an important finding of these within talker effects of experiment II is that the MT segment did provide additional cues to nasal place of articulation for both female and male adult talkers when compared to murmur and transition stimuli of equal duration. This perceptual enhancement was much greater for adult male talkers, but perceptual improvement did occur for both adult groups. Thus for adult speaker groups, these findings for the VC syllable support the "discontinuity hypothesis." However, the MT segment did not provide perceptual improvement over both of the 50M and 50T segments for any of the child speaker groups, again manifesting the developmental nature of the spectral discontinuity property.

3. Comparison of CV and VC productions

In this investigation, the perception of nasals from murmur, transition, and murmur+transition properties was determined for the same speakers' productions of both CV and VC syllables. Examination of speaker group and talker effects clearly revealed differences in the identification of place of articulation of nasals for CV and VC syllables. To provide further comparison between the results for the CV and VC productions, the perceptual data for each syllable type were collapsed across talkers within each speaker group.

a. Murmur cue. There were substantial differences in the identification of nasals from the murmur in VC and CV syllables across duration and this difference was particularly prominent for adult talkers. An analysis of variance revealed significant main effects for syllable type [$F(1,18)=37.33$; $p<0.01$], speaker group [$F(4,72)=67.59$; $p<0.01$], and duration [$F(2,36)=48.14$; $p<0.01$]. There were significant first-order interactions of speaker group × syllable type [$F(4,72)=10.36$; $p<0.01$], duration × syllable type [$F(2,36)=18.99$; $p<0.01$], and speaker group × duration [$F(8,144)=5.68$; $p<0.01$] and a significant second-order interaction of speaker group × duration × syllable type [$F(8,144)=3.92$; $p<0.01$]. According to *posthoc* tests, identification of place of articulation of nasals from the murmur was significantly more accurate from CV syllables than VC syllables for all speaker groups.[12] As shown in Fig. 7 (panel a), consistently more accurate identification of the nasal occurs for adult CV syllables than VC syllables. The same general trend was found for child speaker groups, but the magnitude of the differences was smaller for child speech than adult speech.

b. Transition cue. The results for the transition stimuli are illustrated in Fig. 7 (panel b). These results show that identification of the nasal from the transition cue was generally more accurate in CV syllables than in VC syllables for both adult and child speaker groups. There were significant main effects for syllable type [$F(1,18)=34.21$; $p<0.01$], speaker group [$F(4,72)=26.68$; $p<0.01$], and duration [$F(2,36)=27.06$; $p<0.01$]. Significant first-order interactions were found for speaker group × syllable type [$F(4,72)=19.05$; $p<0.01$] and duration × syllable type [$F(2,36)=74.40$; $p<0.01$]. There was a significant second-order interaction of speaker group × duration × syllable type [$F(8,144)=6.52$; $p<0.01$]. According to *posthoc* tests, nasal identification was significantly more accurate from transitions produced in CV syllables than from transitions produced in VC syllables for the 3- and 5-year-old talkers and for the adult female talkers (FAD).[13] Also illustrated in Fig. 7, identification of nasals from VC and CV transitions was nearly identical at the 25-ms transition (25T) for both adult and child talkers. For all transition duration segments, identification of nasals was more accurate from adult talkers than from child talkers. *Posthoc* tests indicated that identification accuracy was significantly higher from nasals produced by both FAD and MAD speaker groups in CV syllables than for any of the child speaker groups. In contrast, identification was significantly more accurate from nasals produced by only the MAD speaker group in VC syllables than from nasals produced by any of the child speaker groups.[14]

c. Murmur+transition cue. The findings for the murmur+transition stimuli are illustrated in Fig. 7 (panel c). As illustrated in Fig. 5, there were substantive differences in

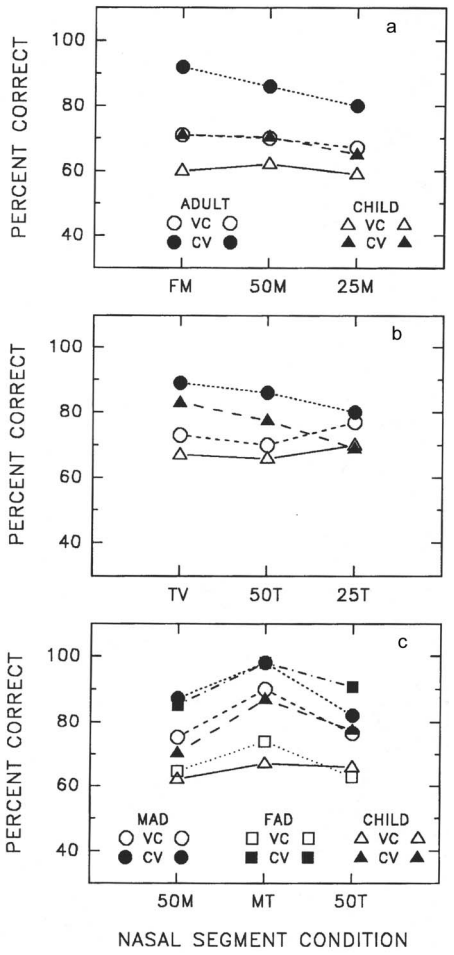


FIG. 7. Mean percent correct identification of nasal consonants from the full murmur (FM), 50-ms murmur (50M), and 25-ms murmur (25M) stimuli (panel a), from the full transition and vowel (TV), 50-ms transition (50T), and 25-ms (25T) transition stimuli (panel b), and from the 50-ms murmur (50M), 25-ms murmur+25-ms transition (MT), and 50-ms transition (50T) stimuli (panel c) collapsed across talkers within a speaker group (FAD: female adults; MAD: male adults) for VC and CV syllables.

the perception of nasal place of articulation from stimuli produced by FAD and MAD speaker groups. To show these differences for the MT segment more clearly, FAD and MAD groups are illustrated separately for this property in Fig. 7 (panel c). An analysis of variance revealed significant main effects of syllable type $[F(1,18)=241.68;\ p<0.01]$ and speaker group $[F(4,72)=86.38;\ p<0.01]$. There was a significant speaker group $\times$ syllable type interaction $[F(4,72)=10.29;\ p<0.01]$. For the MT segment and across speaker groups, identification was significantly more accurate for nasals produced in CV syllables than for nasals produced in VC syllables.[15] As illustrated in Fig. 7 (panel c), there was not a significant difference in the perception of the MT segment from CV syllables produced by male and female adult talkers, but there was a significant difference in the perception of the MT segment between these speaker groups for the VC syllable. Nevertheless, identification of nasals produced by both MAD and FAD groups was significantly more accurate than identification of nasals produced by 3- and 5-year-old speaker groups. Identification of nasals produced by the MAD speaker group was also significantly more accurate than nasals produced by the 7-year-old group in the VC syllable.[16]

In summary, a main finding of experiment II was the significant speaker group effect obtained for the murmur +transition segment in VC syllables. As expected, identification of the nasal from the VC MT segment was more accurate from adult speaker groups than from child speaker groups. An unexpected result in the perceptual identification of VC syllables was the significantly more accurate identification of the nasal from the MT segment produced by the MAD group than the MT segment produced by the FAD group. No such difference was observed for the CV syllables. In addition, the MT segment resulted in accurate and stable perception across talkers when segmented from CV syllables, but resulted in substantial perceptual variability when extracted from VC syllables.

IV. THE ACOUSTIC BASIS OF THE [m]-[n] DISTINCTION IN CV AND VC SYLLABLES: PRELIMINARY FINDINGS

The perceptual findings obtained in this study clearly indicate that the MT segment contains important cues for the perception of the [m] versus [n] distinction. The perceptual results also suggest that there are developmental and syllable effects in the perception of nasal place of articulation from the MT segment. For example, in CV syllables, perception of nasal place of articulation was not significantly different for the MT segment and whole syllable in adult speech, whereas for child speech these cues were significantly different, with the whole syllable carrying greater perceptual weight than the MT segment. In regard to syllable type, perceptual cues to place of articulation of nasals were more robust in the CV than the VC syllable across speaker groups.

In previous research (Kurowski and Blumstein, 1987), comparison of the changes in spectral patterns from the murmur to the transition showed that for labials there was a greater change in energy in low frequencies relative to that in high frequencies, whereas for alveolars there was a greater change in energy from the murmur to the transition in high frequencies relative to that in low frequencies. Based on the perceptual findings in the current research, two predictions would be that this relative amplitude spectral cue would differentiate more the [m] versus [n] distinction in adult speech than in children's speech, and differentiate more the [m] versus [n] distinction in CV syllables than in VC syllables.

As a preliminary assessment of these predictions, high- and low-frequency relative amplitude were examined in the 25M (25-ms murmur; see Figs. 1 and 4) and 25T (25-ms transition; see Figs. 1 and 4) stimulus segments produced by the adult male, adult female, and 3-year-old child speaker groups in experiments I and II. The stimulus segments were the same ones used in the perceptual experiments I and II. Measures were obtained for [m] and [n] 25-ms segments for CV and VC syllables in the context of the [æ] and [*a*] vowels. Low vowels were analyzed, because these vowels most consistently contained the relative amplitude property differentiating [m] and [n] in previous research (Kurowski and Blumstein, 1987).

TABLE I. Relative amplitudes (in dB) of nasal place of articulation for syllable type (CV and VC) for the 3 year old, female adult (FAD), and male adult (MAD) speaker groups.

	[m]		[n]	
	CV	VC	CV	VC
3 year-old	4.5125	1.3000	8.5875	4.0625
FAD	−0.0667	−0.7917	6.2500	6.2583
MAD	0.2167	−5.7333	8.6917	4.0250

Three participants each in the adult male (MAD) and adult female (FAD) groups and two participants in the 3-year-old child group served as talkers. Because one of the original 3-year-old talkers did not produce murmurs in the VC syllable condition, that talker was not included in the acoustic analyses. There were 12 stimulus tokens (2 vowels × 2 productions × 3 talkers) analyzed for each of the four syllables ([m]—CV, [n]—CV, [m]—VC, [n]—VC) for a total of 48 stimulus tokens each for the MAD and FAD groups. For the 3-year-old child group, there were 8 stimulus tokens (2 vowels × 2 productions × 2 talkers) analyzed for each of the four syllables for a total of 32 stimulus tokens. Thus, a total of 128 stimulus tokens were analyzed for the relative amplitude variation in the [m] and [n] nasals.

Waveform analysis was performed using commercially available sound editing software (Adobe Audition version 1.5). A fast Fourier transform analysis (FFT: 1024 points, applying a Hamming window) was performed on the 25M and 25T segments. Using the amplitude information from the FFT analysis, intensities over a range of low (390–780 Hz) and high (1270–2320 Hz) frequencies were summed to determine the overall level in both frequency bands. These frequency regions were selected to be similar to those analyzed by Kurowski and Blumstein (1987). The relative change in amplitude (in dB) for the low- and high-frequency regions was determined by subtracting the level obtained in the murmur segments (25M) from the level obtained in the transition segments (25T) in each utterance. The results for place of articulation and speaker group are shown in Table I. The values in Table I represent a difference score derived by subtracting the relative change in amplitude (in dB) in the low-frequency band from the relative change in amplitude (in dB) in the high-frequency band. Based on previous work (Kurowski and Blumstein, 1987) and on current hypotheses, this value should be negative for [m] and positive for [n]. As shown in Table I, the findings for the FAD and MAD speaker groups are consistent with the above directional effects, whereas for the 3-year-old speaker group productions both [m] and [n] are positive. The general directional effects for [m] and [n] are consistent across speaker groups, but the findings are less robust for the children, which is consistent with the perceptual findings in experiments I and II.

The results of these acoustic analyses support the prediction that the relative amplitude spectral cue would differentiate more the [m] versus [n] distinction in adult speech than in child speech, and that it would differentiate place of articulation better for CV syllables than for VC syllables. An analysis of variance revealed significant main effects of syllable type [$F(1,29)=12.69$; $p<0.01$] and place of articulation [$F(1,29)=54.83$; $p<0.01$] and a significant place × speaker group interaction [$F(2,29)=3.38$; $p<0.05$]. The syllable type × speaker group interaction was not significant. Thus, the significant main effect of syllable type indicates that the relative amplitude cue is represented differently in CV and VC syllables. *Posthoc* tests revealed highly significant differences in the relative amplitude property of [m] and [n] for the FAD, [$F(1,29)=8.98$; $p<0.01$] and MAD [$F(1,29)=23.29$; $p<0.01$] productions, whereas these same differences were only marginally significant [$F(1,29)=4.05$; $p<0.06$] for the 3-year-old speaker group productions.

V. GENERAL DISCUSSION

The purpose of this research was to provide answers to the following questions: (1) Do spectral discontinuities provide cues for accurate identification of place of articulation of nasal consonants across speaker group, talker, and syllable position as predicted by the "discontinuity hypothesis?" (2) Is the CV syllable perceptually more robust than the VC syllable? (3) Are there developmental speaker group and talker differences reflected in adults' perception of murmur and transition segments as cues to place of articulation of nasals?

Based on the current findings, it is clear that in CV syllables the spectral discontinuities of the MT segment provided important cues for accurate identification of place of articulation of nasal consonants across speaker group, talker, and syllable position as predicted by the "discontinuity hypothesis." The talker effect within speaker groups was not eliminated statistically for all these groups, but the MT segment consistently provided the most important information for place of articulation across all groups. The accuracy in identification of nasals produced by all speaker groups was higher for the MT segment than for all murmur or transition segments, even when equated for duration (50 ms). These findings for the MT segment in CV syllables provide support for the perceptual importance of regions with spectral discontinuity between consonant and vowel segments relative to other syllable segments. Acoustic analyses supported the results of the perceptual analyses. Spectrographic analyses and measures of relative amplitude change showed abrupt changes during the brief time interval from the murmur to the vowel transition. Stevens (1985) has suggested that these acoustic discontinuities at the boundaries between two sounds not only serve as a marker between sounds, i.e. segmentation, but also contain important information on the phonetic features of segments. The current findings for the CV MT segment support Stevens' view on the importance of spectral discontinuities at acoustic boundaries as carriers of phonetic feature information.

For VC syllables, the acoustic boundary between the vowel and the nasal consonant provided more useful phonetic information than the separate transition and murmur cues of equivalent duration only for adult male and female speakers. In these speakers, the listeners' accuracy in identification of the VC MT segment was lower than their respec-

tive accuracy in identification of the CV MT segment. A comparison of Figs. 3 (panel c) and 6 (panel c) clearly illustrates these differences. For example, there is nearly a 40% difference between CV and VC syllables for FAD speaker T2. The perceptual differences are supported by acoustic analyses showing a difference in the relative spectral change in VC syllables compared to CV syllables. Similar preliminary acoustic differences for CV and VC syllables were reported by Kurowski and Blumstein (1987). The "discontinuity hypothesis" predicts that spectral change at consonant-vowel boundaries contains important information on the phonetic features of segments independent of syllable shape. The current findings for both CV and VC syllables produced by adult speakers support the "discontinuity hypothesis," as presented in our first question. However, the relative perceptual cue value of spectral change and therefore support for the "discontinuity hypothesis" was greater in CV syllables than VC syllables.

Although the perceptual cues to place of articulation of nasals in VC syllables were relatively weak in comparison to those in CV syllables, it is hypothesized that acoustic information carried by the CV MT segments may serve to enhance feature perception in other syllable types such as the VC syllable through a form of feature training and learning early in a child's sound development of place of articulation of nasals. The MT segment produced by adult speakers in CV syllables generally contains more robust cues to nasal identification than any other segment across syllable shape. The process of learning sound features such as place of articulation from CV syllables is supported in sound development where the CV syllable appears first in babbling, ensuring that perceptual exposure to this production event will be frequent and common across the important developmental period. Even though the VC syllable segments were not as perceptually distinct as the CV syllable segments, it is possible that perceptual learning of those cues in the CV murmur and transition that differentiate place of articulation will form the basis for relatively high identification of nasals from whole VC syllables as found in the current study for a range of speaker ages and talkers within speaker groups. This view on learning place of articulation of nasals in one syllable type from the exposure to those properties in another syllable shape is consistent with the view that nasal place of articulation is most accurately classified acoustically by combining the separate murmur and vowel spectra (Harrington, 1994). In this view, the fact that VC syllables lack a spectral discontinuity for many talkers is not critical. What is critical is that the separate murmurs and transitions contain important cues for place of articulation of nasals. As the current research shows, these cues contain significant information for place of articulation of the nasal particularly for adults' CV productions. In addition, combining the murmur and vowel spectra provided the most important cues to nasal identification for both CV and VC syllables, particularly in adult speech. Because adults serve as acoustic models for children learning the speech features of a language, the MT integrated cue would appear to have special linguistic status in this acquisition process, perhaps across syllable shapes. The importance of this integrated cue is further highlighted for the CV condition where the identification of nasals from the FAD and MAD MT segment was as accurate as identification of nasals from their whole syllable productions.

The findings of this research clearly show that the CV syllable is perceptually distinctive compared to the VC syllable, as addressed in our second question. The evidence for this conclusion was strong in that identification of place of articulation of nasals was significantly more accurate from the CV syllable than from the VC syllable across the segments of murmur, transition, and murmur+transition.

In regard to the third question on the development of cue distribution for place of articulation, the results of experiment I for CV syllables provide corroborating evidence of previous research employing a single talker per speaker group (Ohde, 1994). In particular, these studies show that the identification of nasals from the murmur and the integrated murmur+transition segment is significantly more accurate for adult speech than for child speech. A comparison of findings for the identification of nasals from the murmur and transition segments revealed differences between adult and child speaker groups. The results for adult speaker groups show that identification of place of articulation from murmurs and transitions is relatively equivalent, thus supporting previous findings of Kurowski and Blumstein (1984). In children's production of CV syllables, on the other hand, adults apparently perceive place of articulation for nasal consonants more accurately from vowel transitions than from murmurs. The dominance of coarticulated transition (dynamic) properties over segmental murmur (static) properties suggests that dynamic features may also have greater prominence in children's speech production. Interestingly, research has also shown that children place greater weight on dynamic acoustic properties than on static acoustic properties in their perception of consonant place of articulation (Nittrouer, 1992; Ohde and Haley, 1997; Bourland Hicks and Ohde, 2005).

The findings for the VC syllable were different from those for the CV syllable as related to development. For all speaker groups, listeners were more accurate in identifying place of articulation of nasals from the CV MT condition than from the 50M or 50T segments [see Fig. 2 (panel c)]. Listeners' identification was nearly 100% accurate from adult male and female speaker groups, with 7-year-old and 5- and 3-year-old groups in the 90% and 80% ranges, respectively. However, in the VC MT condition, identification of the nasal was more accurate from the MT segment than from the 50M or 50T segments for adult speaker groups only [see Fig. 5 (panel c)]. Thus, the systematic developmental effects found for the CV murmur and transition conditions were not observed in the VC conditions. Also, identification of the female adult place of articulation conditions and the child conditions was somewhat similar for the murmur and transition isolated segments [see Fig. 5 (panels a and b)]. It is quite possible that a combination of articulatory factors such as a gradual versus an abrupt opening of the velopharyngeal port, and source factors such as a relatively high f_0 and breathiness for female adults compared to male adults, contributed to the similarity in perception of place of articulation of adult female speech and child speech in the VC conditions. Research has shown that perception is poorer for vowels with high f_0's

compared to low f_0's (Ryalls and Lieberman, 1982), and that female voices contain greater breathiness than male voices (Klatt and Klatt, 1990). As well known, adult female and child voices also have higher f_0's than male adult voices (Perry *et al.*, 2001). These articulatory and source factors must uniquely interact with VC productions, since identification of nasals from the CV MT stimulus produced by adult female and male speakers were identical and near 100%.

The results of experiment I for CV syllables also revealed significant differences in the perception of place of articulation for murmur and transition cues across talkers within a speaker group. The talker effect was significant for all speaker groups except the 7-year old. For some segments, variation in perception of place of articulation varied by over 20% for talkers within a speaker group. The selection process for including talkers within a speaker group was intended to represent a range of clarity of production for the CV syllables, and the variation in perception due to talker did appear in part to relate to this selection process. Although the talkers in this research were native speakers of English whose whole syllables were highly identifiable as their intended target productions, the reduced talker variability as reflected in perception of nasals in CV syllables was due to the perceptual prominence of the spectral discontinuity in the MT segments of CV syllables.

In summary, the findings of this study revealed that the cues to the identification of place of articulation of nasals in VC syllables were impoverished compared to the same properties in CV syllables. Moreover, the combination of murmur and transition in the MT segment did not stabilize perceptual variability across speakers in the VC syllable as it did in the CV syllable. Developmental effects were observed in VC syllables, as reflected in the more perceptually prominent adult MT segment than child MT segment. The importance of these findings is that they provide insight into phonological development and perceptual distinctiveness as a function of syllable type from a perceptual perspective. Although the majority of previous developmental research has examined only CV syllables, a complete understanding of phonological development must include other syllable shapes in both perception and production paradigms. In comparing the findings for CV and VC syllables, the integrated murmur+transition property reflected the greatest difference between these productions, with the CV syllable characterized by high phonetic content at the point of spectral discontinuity. It is hypothesized that since regions of spectral discontinuity are particularly strong in CV syllables, these phonetically high content, perceptually stabilizing, and distinctive properties may in part be the basis of the early CV acquisition in children, the training source for the place of articulation of nasals in other syllable shapes such as the VC, and the universality of the CV syllable shape across languages (Ohde, 1994; Ohde and Haley, 1992).

VI. CONCLUSIONS

A primary goal of the current research was to test the "discontinuity hypothesis" (Stevens, 1985) as a viable description of the perception of nasal consonants in CV and VC syllables. The following conclusions were supported by this research:

(1) As predicted by the "discontinuity hypothesis," the stimulus segment containing the spectral discontinuity provided the strongest perceptual cues to place of articulation of nasals in both CV and VC syllables for adult speaker groups, but more so in the former than latter syllable shape.
(2) The stimulus segment containing the spectral discontinuity was least variable as a perceptual cue in only the CV syllable.
(3) Place of nasal articulation was more accurately identified in the CV syllable than in the VC syllable across speaker group and talker.
(4) Both talker and developmental differences were observed for the murmur and transition properties as evidenced in the perception of these cues to place of articulation of nasals.
(5) The relative amplitude change from the murmur to vowel transition in low- versus high-frequency regions more strongly differentiated place of articulation in adult speaker groups than in the 3-year-old speaker group.

ACKNOWLEDGMENTS

This research was supported in part by NIH Grant Nos. DC00464 and DC00523-08. We thank Ben Hornsby for assistance with acoustic analyses, Courtney Jacks for assistance with data analyses, and we extend our appreciation to Randy Diehl and two anonymous reviewers for their excellent reviews of this paper. Finally, we thank the adults and the children and their parents who participated in this research, without whose cooperation, help, and patience such study could not have been completed.

[1]3-year: $[F(2,18)=10.14;\ p<0.01]$; 7-year: $[F(2,18)=7.87;\ p<0.01]$; FAD: $[F(2,18)=35.89;\ p<0.01]$; MAD: $[F(2,18)=15.06;\ p<0.01]$.
[2]FM: $[F(4,36)=69.84;\ p<0.01]$; 50M: $[F(4,36)=35.42;\ p<0.01]$; 25M: $[F(4,36)=21.82;\ p<0.01]$.
[3]3-year: $[F(2,18)=34.91;\ p<0.01]$; 5-year: $[F(2,18)=59.52;\ p<0.01]$; 7-year: $[F(2,18)=24.43;\ p<0.01]$; FAD: $[F(2,18)=28.28;\ p<0.01]$; MAD: $[F(2,18)=6.34;\ p<0.01]$.
[4]TV: $[F(4,36)=11.53;\ p<0.01]$; 50T: $[F(4,36)=12.93;\ p<0.01]$; 25T: $[F(4,36)=13.97;\ p<0.01]$.
[5]3-year: $[F(4,36)=81.66;\ p<0.01]$; 5-year: $[F(4,36)=87.49;\ p<0.01]$; 7-year: $[F(4,36)=75.84;\ p<0.01]$; FAD: $[F(4,36)=43.73;\ p<0.01]$; MAD: $[F(4,36)=32.39;\ p<0.01]$.
[6]3-year: $[F(4,36)=84.81;\ p<0.01]$; 5-year: $[F(4,36)=78.92;\ p<0.01]$; 7-year: $[F(4,36)=93.16;\ p<0.01]$; FAD: $[F(4,36)=40.09;\ p<0.01]$; MAD: $[F(4,36)=40.07;\ p<0.01]$.
[7]3-year: T3 $[F(2,18)=9.06;\ p<0.01]$; 5-year: T2 $[F(2,18)=11.24;\ p<0.01]$; T3 $[F(2,18)=38.22;\ p<0.01]$. Identification of place of articulation was significantly more accurate from the MT segment than either the 50M or 50T segments for all three of the 7-year-old talkers. FAD: T2 $[F(2,18)=8.28;\ p<0.01]$; T3 $[F(2,18)=12.52;\ p<0.01]$; MAD: T2 $[F(2,18)=4.12;\ p<0.05]$; T3 $[F(2,18)=18.90;\ p<0.01]$.
[8]3-year: $[F(2,18)=6.68;\ p<0.01]$; FAD: $[F(2,18)=15.92;\ p<0.01]$.
[9]3-year: $[F(4,36)=58.63;\ p<0.01]$; 5-year: $[F(4,36)=157.75;\ p<0.01]$; 7-year: $[F(4,36)=97.04;\ p<0.01]$; FAD: $[F(4,36)=120.44;\ p<0.01]$; MAD: $[F(4,36)=93.89;\ p<0.01]$.
[10]FAD: $[F(4,36)=73.01;\ p<0.01]$; MAD: $[F(4,36)=52.27;\ p<0.01]$.
[11]FAD: T1 $[F(2,18)=17.57;\ p<0.01]$; T3 $[F(2,18)=23.18;\ p<0.01]$; MAD: T1 $[F(2,18)=23.97;\ p<0.01]$; T2 $[F(2,18)=11.55;\ p<0.01]$.

[12]3-year: $[F(1,18)=11.64;\ p<0.01]$; 5-year: $[F(1,18)=6.33;\ p<0.05]$; 7-year: $[F(1,18)=28.53;\ p<0.01]$; FAD: $[F(1,18)=61.63;\ p<0.01]$; MAD: $[F(1,18)=15.52;\ p<0.01]$.
[13]3-year: $[F(1,18)=45.73;\ p<0.01]$; 5-year: $[F(1,18)=27.34;\ p<0.01]$; FAD: $[F(1,18)=76.93;\ p<0.01]$.
[14]CV age effects: $[F(4,72)=25.73;\ p<0.01]$; VC age effects: $[F(4,72)=20.00;\ p<0.01]$.
[15]3-year: $[F(1,18)=153.24;\ p<0.01]$; 5-year: $[F(1,18)=60.11;\ p<0.01]$; 7-year: $[F(1,18)=59.91;\ p<0.01]$; FAD: $[F(1,18)=161.82;\ p<0.01]$; MAD: $[F(1,18)=28.01;\ p<0.01]$.
[16]CV age effects: $[F(4,72)=47.10;\ p<0.01]$; VC age effects: $[F(4,72)=49.57;\ p<0.01]$.

Ali, L., Gallagher, T., Goldstein, J., and Daniloff, R. (**1971**). "Perception of coarticulated nasality," J. Acoust. Soc. Am. **49**, 538–540.

Bankson, N. (**1977**). *The Bankson Language Screening Test*. (University Park, Baltimore, MD).

Bourland Hicks, C., and Ohde, R. N. (**2005**). "Developmental role of static, dynamic, and contextual cues in speech perception," J. Speech Lang. Hear. Res. **48**, 960–974.

Delgutte, B. (**1980**). "Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers," J. Acoust. Soc. Am. **68**, 843–857.

Delgutte, B., and Kiang, N. Y.-S. (**1984**). "Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics," J. Acoust. Soc. Am. **75**, 897–907.

Dyson, A. T. (**1988**). "Phonetic inventories of 2- and 3-year-old children," J. Speech Hear Disord. **53**, 89–93.

Fluharty, N. (**1974**). "The design and standardization of a speech and language screening test for use with preschool children," J. Speech Hear Disord. **39**, 75–88.

Fluharty, N. (1978). *Fluharty Preschool Speech and Language Screening Test*, Teaching Resources Corporation, Hingham, MA.

Fudala, J. B. (1970). *Arizona Articulation Proficiency Scale-Revised*, Western Psychological Services, Los Angeles, CA.

Halle, M., and Stevens, K. N. (**1991**). "Knowledge of language and the sounds of speech," in *Music, Language, Speech and Brain*, edited by J. Sunberg, L. Nord, and R. Carlson (MacMillan, London), pp. 1–19.

Hamby, M. J. (2000). "The development of coarticulatory and segmental properties in nasal+vowel syllables," unpublished master's thesis, Vanderbilt University, Nashville, TN.

Harrington, J. (**1994**). "The contribution of the murmur and vowel to the place of articulation distinction in nasal consonants," J. Acoust. Soc. Am. **96**, 19–32.

Kent, R. D., and Murray, A. D. (**1982**). "Acoustic features of infant vocalic utterances at 3, 6, and 9 months," J. Acoust. Soc. Am. **72**, 353–365.

Klatt, D. H. (**1986**). "The problem of variability in speech recognition and models of speech perception," in *Invariance and Variability in Speech Processes*, edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 300–324.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.

Kurowski, K., and Blumstein, S. E. (**1984**). "Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants," J. Acoust. Soc. Am. **76**, 383–390.

Kurowski, K., and Blumstein, S. E. (**1987**). "Acoustic properties for place of articulation in nasal consonants," J. Acoust. Soc. Am. **81**, 1917–1927.

Kurowski, K. M., and Blumstein, S. E. (**1993**). "Acoustic properties for the perception of nasal consonants," in *Nasals, Nasalization, and the Velum*, edited by M. K. Huffman and R. A. Krakow (Academic, New York), pp. 197–222.

Manuel, S. Y. (**1991**). "Some phonetic bases for the relative malleability of syllable-final versus syllable-initial consonants," in Proceedings of the 12th International Congress of Phonetics Sciences, Vol. **5**, pp. 118–121.

Nittrouer, S. (**1992**). "Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries," J. Phonetics **20**, 351–382.

Ohala, J. J. (**1990**). "The phonetics and phonology of aspects of assimilation," in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. Beckman (Cambridge, U.P., Cambridge), pp. 258–275.

Ohde, R. N. (**1985**). "Fundamental frequency correlates of stop consonant voicing and vowel quality in the speech of preadolescent children," J. Acoust. Soc. Am. **78**, 1554–1561.

Ohde, R. N. (**1994**). "The development of cues to the perception of the /m/-/n/ distinction in CV syllables," J. Acoust. Soc. Am. **96**, 675–686.

Ohde, R. N., and Haley, K. (**1992**). "The developmental role of interspeaker variability in the perception of the /m/-/n/ distinction of CV syllables," J. Acoust. Soc. Am. **92**(Pt. 2), 2463(A).

Ohde, R. N., and Haley, K. L. (**1997**). "Stop-consonant and vowel perception in 3- and 4-year-old children," J. Acoust. Soc. Am. **102**, 3711–3722.

Ohde, R. N., and Ochs, M. T. (**1996**). "The effect of segment duration on the perceptual integration of nasals for adult and child speech," J. Acoust. Soc. Am. **100**, 2486–2499.

Ohde, R. N., and Perry, A. M. (**1994**). "The role of short-term and long-term auditory storage in processing spectral relations for adult and child speech," J. Acoust. Soc. Am. **96**, 1303–1313.

Ohde, R. N., and Sharf, D. J. (**1992**). *Phonetic Analysis of Normal and Abnormal Speech* (Macmillan, Columbus, OH).

Parnell, M. M., Amerman, J. D., and LaRiviere, C. W. (**1978**). "Influence of speaker age on perceptual cue distribution," J. Phonetics **6**, 275–282.

Perry, T. L., Ohde, R. N., and Ashmead, D. H. (**2001**). "The acoustic bases for gender identification from children's voices," J. Acoust. Soc. Am. **109**, 2988–2998.

Polgar, G., and Weng, T. (**1979**). "The functional development of the respiratory system: From the period of gestation to adulthood," Am. Rev. Respir. Dis. **120**, 625–695.

Redford, M. A., and Diehl, R. L. (**1999**). "The relative perceptual distinctiveness of initial and final consonants in CVC syllables," J. Acoust. Soc. Am. **106**, 1555–1565.

Repp, B. H. (**1986**). "Perception of the [m]-[n] distinction in CV syllables," J. Acoust. Soc. Am. **79**, 1987–1999.

Repp, B. H. (**1987**). "On the possible role of auditory short-term adaptation in perception of the prevocalic [m]-[n] contrast," J. Acoust. Soc. Am. **82**, 1525–1538.

Repp, B. H., and Svastikula, K. (**1988**). "Perception of the [m]-[n] distinction in VC syllables," J. Acoust. Soc. Am. **83**, 237–247.

Ryalls, J. H., and Lieberman, P. (**1982**). "Fundamental frequency and vowel perception," J. Acoust. Soc. Am. **72**, 1631–1634.

Seitz, P. F., McCormick, M. M., Watson, I. M. C., and Bladon, R. A. (**1990**). "Relational spectral features for place of articulation in nasal consonants," J. Acoust. Soc. Am. **87**, 351–358.

Stevens, K. N. (**1985**). "Evidence for the role of acoustic boundaries in the perception of speech sounds," in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, edited by V. A. Fromkin (Academic, London), pp. 243–255.

Stevens, K. N. (**1989**). "On the quantal nature of speech," J. Phonetics **17**, 3–46.

Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT, Cambridge, MA).

Stevens, K. N. (**2002**). "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**, 1872–1891.

Stoel-Gammon, C. (**1985**). "Phonetic inventories, 15–24 months: A longitudinal study," J. Speech Hear. Res. **28**, 505–512.

Sussman, H. M., Bessell, N., Dalston, E., and Majors, T. (**1997**). "An investigation of stop place of articulation as a function of syllable position: A locus equation perspective," J. Acoust. Soc. Am. **101**, 2826–2838.

Wright, R. (**2001**). "Perceptual cues in contrast maintenance," in *The Role of Speech Perception in Phonology*, edited by E. Humes and K. Johnson (Academic, New York), pp. 251–277.

Extrinsic context affects perceptual normalization of lexical tone

Alexander L. Francis[a)]
Department of Speech, Language and Hearing Sciences, Purdue University, Heavilon Hall, 500 Oval Drive, West Lafayette, Indiana 47907

Valter Ciocca, Natalie King Yu Wong, Wilson Ho Yin Leung, and Phoebe Cheuk Yan Chu
Division of Speech and Hearing Sciences, Faculty of Education, 5th Floor, Prince Philip Dental Hospital, 34 Hospital Road, Hong Kong

(Received 28 April 2005; revised 7 October 2005; accepted 11 October 2005)

The present study explores the use of extrinsic context in perceptual normalization for the purpose of identifying lexical tones in Cantonese. In each of four experiments, listeners were presented with a target word embedded in a semantically neutral sentential context. The target word was produced with a mid level tone and it was never modified throughout the study, but on any given trial the fundamental frequency of part or all of the context sentence was raised or lowered to varying degrees. The effect of perceptual normalization of tone was quantified as the proportion of non-mid level responses given in F0-shifted contexts. Results showed that listeners' tonal judgments (i) were proportional to the degree of frequency shift, (ii) were not affected by non-pitch-related differences in talker, (iii) and were affected by the frequency of both the preceding and following context, although (iv) following context affected tonal decisions more strongly than did preceding context. These findings suggest that perceptual normalization of lexical tone may involve a "moving window" or "running average" type of mechanism, that selectively weights more recent pitch information over older information, but does not depend on the perception of a single voice. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2149768]

PACS number(s): 43.71.Bp, 43.71.−k, 43.71.An, 43.71.Es, 43.71.Hw [ARB] Pages: 1712–1726

I. INTRODUCTION

Listeners may perceive the same acoustic pattern as different phonemes depending on the phonetic context it appears in (Liberman *et al.*, 1967), the speaking rate at which it appears to be produced (Verbrugge *et al.*, 1976), and the talker that is perceived to have produced it (Ladefoged and Broadbent, 1957). Such context-dependent processing of speech can be referred to as (perceptual) normalization, or more specifically according to the nature of the contextual information being used (e.g., phonetic normalization, speaking rate normalization, talker normalization). Note that this use of the term normalization may be distinguished from another use referring to the physical transformation of a signal to reduce between-token variability (e.g., peak amplitude normalization, or acoustic normalization more generally). The two uses of the term are historically related, in the sense that early conceptualizations of perceptual normalization assumed that listeners were mentally transforming incoming signals to derive context-independent representations in the same way that an engineer might use acoustic normalization to transform a physical signal to eliminate context-specific variability (see Johnson, 1997 for discussion). The assumption that perceptual normalization must result in a context-independent mental representation of the signal is not, however, a necessary one for the purposes of the present discussion (Johnson, 1997; Nusbaum and Magnuson, 1997). Here we will refer to talker normalization or tone normalization in keeping with the established literature on the topic, though we use these terms to refer simply to the process by which listeners understand speech (and lexical tones in particular) in a talker-dependent manner.

Talker normalization has been demonstrated for vowels (Ladefoged and Broadbent, 1957; Nearey, 1989), consonants (Johnson, 1991), and lexical tone (Jongman and Moore, 2000; Moore and Jongman, 1997; Wong, 1998; Wong and Diehl, 2003). For example, Ladefoged and Broadbent (1957) showed that a synthetic vowel in a "b_t" context was perceived as "bit" in isolation but as "bet" when preceded by a precursor sentence (Please say what this word is:) that was synthesized to have a generally lower first formant (F1) frequency. In other words, listeners interpreted the "bit" token as having a comparatively higher F1 when it was presented in the context of a sentence with overall lower F1 values.

The above-noted examples show that perceptual normalization, particularly talker normalization, derives at least in part from information provided by "extrinsic context" (or "extrinsic information")—speech that does not constitute part of the syllable or phoneme to be identified. Such extrinsic information seems to be particularly significant in the case of perception of lexical tones. In a lexical tone language, meaning can be distinguished according to the suprasegmental feature of tone alone. In Cantonese, the primary physical correlate of tone is fundamental frequency (f0) (Fok Chan, 1974; Vance, 1976). Thus, two syllables with different f0 contours may have different meanings, even when their

[a)]Electronic mail: francisa@purdue.edu

 0001-4966/2006/119(3)/1712/15/$22.50

segmental content is the same. For example, in Cantonese the segmental string /ji/ means "doctor" when produced with a high level tone but it means "two" when produced with a low level tone.[1] What is particularly interesting about the Cantonese tonal system is that it contrasts three level tones (high, mid, and low, or 55, 33, and 22) that differ minimally in terms of their contours (Bauer and Benedict, 1997; Rose, 2000). As we shall see, this system means that talker normalization on the basis of extrinsic context plays a particularly significant role in the accurate perception of Cantonese tones.

Since most research on tone normalization has been conducted using Mandarin listeners and stimuli, it is instructive to first review these studies in order to better understand the phenomenon in question. Standard (Beijing) Mandarin has four tone categories: a high (55) tone (tone 1), a rising (25) tone (tone 2), a dipping (214) tone (tone 3), and a falling (51) tone (tone 4). Although Mandarin tones differ primarily in terms of f0 contour, they can also be affected by their perceived relative height. For example, Leather (1983) showed that the perception of syllables with f0 contours lying toward the middle of a continuum between that of a 25 (tone 2) and 55 (tone 1) tone were interpreted differently depending on the context in which they appeared.

Tone normalization has been studied in Mandarin by Leather (1983), Lin and Wang (1985), Fox and Qi (1990), and Moore and Jongman (1997). Leather (1983) embedded tonally ambiguous tokens in carrier sentences produced by talkers with very different average f0 ranges, and showed that listeners categorized stimuli with identical f0 according to the speaker characteristics provided by the preceding context. Both Lin and Wang (1985) and Fox and Qi (1990) examined the effect of manipulating the f0 of one syllable in two-syllable sequences. Lin and Wang (1985) examined the effect of manipulating the f0 of the second syllable on perception of the tone of the first syllable, while Fox and Qi (1990) examined the effect of manipulating the f0 of the first syllable on perception of the tone of a following syllable. Although the results of Lin and Wang (1985) provided stronger evidence for contextual normalization of tone, Fox and Qi (1990) also found some evidence to support the idea that tone identification is affected by extrinsic context. Taken together, the results of all of these studies strongly suggest that, when the f0 of the target syllable is held unchanged, the perception of its tone can be influenced by the f0 of a neighboring syllable. Moore and Jongman (1997) provided the clearest demonstration of this phenomenon to date. They showed that a given (synthesized) Mandarin Chinese syllable could be identified as having either of two different tones depending on the fundamental frequency (f0) of the preceding sentence.[2] When the context f0 was low, the target was identified as having a mid rising tone (tone 2). When the context f0 was higher, the identical target syllable was identified as having a low falling-rising tone (tone 3). Thus, even though Mandarin tones may be identifiable on the basis of other (non-f0) properties (Fu *et al.*, 1998; Whalen and Xu, 1992), there is a clear and well-established effect of the fundamental frequency properties of preceding speech on the perception of Mandarin lexical tones.

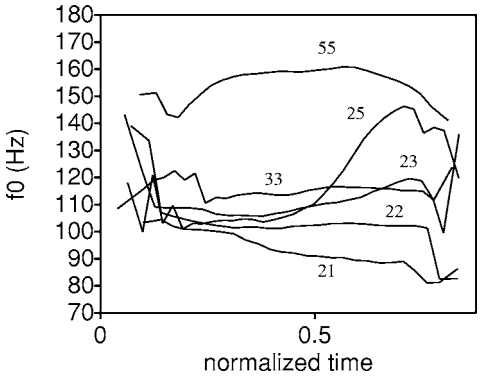


FIG. 1. Fundamental frequency (f0) contours for the six Cantonese tones produced by the speaker on whose speech the stimuli for experiment 1 were modeled.

Unlike Mandarin, Cantonese has three level tones that differ only in terms of their relative (average) f0 (Bauer and Benedict, 1997; Matthews and Yip, 1994; Rose, 2000; Vance, 1976; Wong, 1998; Wong and Diehl, 2003) as well as three contour tones (see Fig. 1). Thus, it is entirely possible that a low level tone produced by a talker with a very high average f0 might easily fall within the range of a high level tone produced by a talker with an overall lower average f0. This makes it imperative that listeners be able to judge tone within the context of the talker's individual f0 characteristics. Research by Wong (1998) and Wong and Diehl (2003) demonstrated the importance of extrinsic context in this process. They showed that, in Cantonese, it is possible to embed a syllable with a mid level tone (33) in a variety of contexts differing only in average f0, and have that tone be interpreted as either high, mid, or low level depending on the perceived f0 of the context. Wong (1998) and Wong and Diehl (2003) made use of this effect and manipulated the average fundamental frequency of a sentence preceding a target syllable produced with a mid level (33) tone. They showed that shifting the frequency of the preceding sentence upward by two semitones caused Cantonese listeners to hear the target as having a low level tone, while shifting the precursor frequency downward by three semitones caused the target to be identified as having a high level tone. Wong (1998) showed that this effect could be obtained not only with a Cantonese precursor, but also with an English precursor produced by the same talker, although the English effect was comparatively smaller. Since the English phonological system does not make use of lexical tones, these results suggested that Cantonese listeners did not need to access information specific to linguistically defined tonal categories in the precursor sentence. Rather, they could perform the task on the basis of nonlinguistic pitch properties alone. Wong (1998) observed that the overall pitch range of the talkers was smaller in English than in Cantonese. He argued that this suggested that Cantonese listeners were basing their tonal judgments on the position of the target syllable's f0 relative to the maximum (or minimum) pitch of the preceding context (the Pitch Range Assessment Model).

A. The Pitch Range Assessment Model (PRAM)

According to the Pitch Range Assessment Model (PRAM), listeners estimate a talker's tonal range on the basis

of the actual pitch ranges they experience in a given utterance. Therefore, a precursor sentence with a wider pitch range (the Cantonese precursor) should provide a more accurate estimation of expected tone productions than does one with a more constrained range (the English precursor). Similar arguments have been made regarding perceptual normalization of vowel spaces (Joos, 1948, p. 61). While these have generally been superseded by more sophisticated models, the reason for this seems to be at least in part because talker-specific vowel identification is quite accurate even in isolated syllables (Verbrugge *et al.*, 1976; see Johnson, 2005 for discussion). As we shall see here, although Cantonese tone identification is possible at better-than-chance levels in isolated syllables (Francis *et al.*, 2003; Wong, 1998; Wong and Diehl, 2003) and even without segmental information (Fok Chan, 1974), there is a substantial effect of extrinsic context, especially compared to that observed by Verbrugge *et al.* (1976) for vowels (see also Francis *et al.* 2003). Thus, a model such as PRAM may be more plausible in the case of tone normalization than it might be for vowel normalization. On the other hand, the PRAM was proposed on the basis of a very small number of experiments, and its predictive power remains limited. Many factors governing the operation of the PRAM remain to be determined.

B. The importance of pitch range

PRAM suggests that listeners will perform best when the context provides the widest possible frequency range for a given talker. Wong's (1998) results support this hypothesis, but could also be explained by listeners' different expectations regarding the role of pitch in English versus Cantonese rather than by any effect of pitch range per se. If bilingual Cantonese and English listeners expect f0 to play a different role in English than in Cantonese, they may rely less on contextual information about f0 when judging Cantonese syllables produced in an English context as compared to when they appear in a Cantonese context.[3] Thus, the apparent effect of talker pitch range may be spurious. Instead of using a relatively complex and highly variable measure such as pitch range over the course of a short, recent utterance to estimate a talker's tonal space, listeners might accomplish tone normalization on the basis of some more basic property of the talker's voice, for example average pitch, or even some non-pitch property of speech, perhaps one relating to overall vocal tract size (as has also been suggested for vowel normalization, see Johnson (2005) for details). To further explore this possibility, in experiment 1 we examined the response patterns of listeners hearing only one language (to eliminate effects of listeners' language-based expectations), and instead artificially varied the pitch range provided in the context.

C. Degree of shift

A second potential problem for the PRAM derives from the way stimuli were constructed. Wong (1998) and Wong and Diehl (2003) showed that a two-semitone upward shift, or a three semitone downward shift, were sufficient to induce tone normalization. These shift values were chosen because Chao (1947) reported that the Cantonese low level tone was approximately two semitones lower than the mid level tone, which was in turn approximately three semitones lower than the high level tone. However, data presented by Rose (2000; personal communication) suggests that the actual difference in modern Hong Kong Cantonese is approximately 4.1 semitones between the high level and mid level tones, and 1.6 semitones between the mid level and low level tones. Thus, it appears that Wong (1998) and Wong and Diehl (2003) may not have provided their listeners with an optimal stimulus configuration, yet their listeners still showed strong evidence for perceptual normalization of tone when stimuli were presented in context. This suggests that listeners do not need the pitch of the target syllable to lie exactly where it would be expected to be on the basis of the rest of the talkers' tone space, but instead accept a relatively wide range of possible frequencies as representative of each lexical tone. To explore this question, in experiment 1 (and subsequently) we also investigate the effect of the degree of contextual shift on tone normalization.

D. Preceding or following context

Third, with the exception of a single study by Lin and Wang (1985), most previous studies have used target syllables in utterance-final position. Thus, in most cases, contextual information was only provided prior to presentation of the target syllable. While such stimuli clearly provide listeners with optimal conditions for performing tone normalization, they cannot provide crucial information regarding the processing mechanisms involved. Most of the results presented thus far are consistent with a purely "feed-forward" model of tone normalization in which (only) previously occurring pitch information is used to estimate the tone category of a given syllable. However, hypothesizing such a model immediately raises the question of how listeners might judge the tone category of the first (or only) syllable in an utterance—a task that is certainly possible, but not well-studied. The results of Lin and Wang (1985) suggest that following context can be sufficient to facilitate tone normalization in the absence of a preceding context, but, because they used only a following context, their results still do not provide much insight into the relative weighting of preceding versus following context in tone normalization. In experiment 3 we explore the relative weighting of preceding and following context in tone normalization.

E. Talker specificity

Finally, previous studies have all implicitly assumed that tone normalization is a phenomenon depending primarily on fundamental frequency, regardless of listeners' perception of who is talking or what language they are speaking. Wong (1998) showed that listeners can accept changes in language, albeit with a slight reduction in the strength of normalization, suggesting that tone normalization is not completely language-specific. Fox and Qi (1990) have also proposed that tone normalization may derive from the operation of a mechanism that is not specific to linguistic processing. They showed that English and Mandarin speakers were equally

influenced by the F0 of a preceding syllable when asked to judge how similar a second syllable was to either Mandarin tone 1 (high level) or tone 2 (rising). Since English listeners were able to perform the task without any linguistic knowledge of the tonal system of Mandarin, Fox and Qi (1990) concluded that listeners depended solely on the acoustic information, i.e., the F0 of the context, to execute tone normalization. Thus, normalization was argued to be an auditory process, rather than a phonetic one.

In further support of the hypothesis that tone normalization does not require listeners to perceive the speech as being produced by a single talker, both Moore and Jongman (1997) and Leather (1983) have shown that tone normalization continues to operate despite (presumed) changes in voice quality within the same sentence. Both studies used synthetic target stimuli embedded within natural context sentences and showed normalization effects, which would suggest that some divergence in voice quality between context sentence and target syllable is acceptable to listeners. However, this aspect of tone normalization has not yet been investigated explicitly. If tone normalization results from the operation of a general auditory process rather than as an aspect of a language-specific process of talker normalization, then we would expect it to be robust even in the face of clearly noticeable changes in linguistic information and talker identity. We explore these factors in the second half of experiment 3 and in experiment 4.

In summary, the PRAM represents a preliminary model of tone normalization. However, tone normalization is not yet well understood, and many aspects remain to be investigated. Here we present the results of four experiments that, taken together, provide more insight into the operation of tone normalization, and suggest ways in which the PRAM can be augmented in order to account for these new phenomena.

II. EXPERIMENT 1

According to the PRAM, lexical tone normalization should be more effective when the extrinsic context provides the fullest possible range of frequency variation for a given talker. That is, if tone normalization depends on the range of frequencies in the context, then normalization should not take place if the context is completely monotone. Alternatively, tone normalization may involve the derivation of expected frequency values or ranges for particular tones based on a more abstract property of the extrinsic context, for example average f0. In this case, tone normalization should function equally well with a sentence with naturally varying frequency as with a sentence with a monotone f0 pattern that has the same average f0 as the natural one. To test this question, two different types of extrinsic context were used, one with a normal range of frequency variation, and one (monotone) in which the f0 of the context phrase was held constant at a value equal to the mean f0 over the other sentence (excluding the target).

Moreover, the original formulation of the PRAM implies that listeners' context-derived expectations of tone category locations should reflect experience with the typical (relative) f0 values of the tones as produced in the ambient language. Thus, for speakers of a language where the high level tone is typically produced at a frequency three semitones higher than the mid level tone, hearing a particular talker would induce a mental representation of that talker's tone space in which the high level tone is expected to be three semitones higher than the mid level tone. However, the degree of precision of such expectations is unknown. Is it sufficient that a token merely be higher in frequency than the expected frequency of a mid level tone in order for that token to be heard as a high level tone, or must the f0 value of the target syllable match the expected f0 of the high level tone category more precisely? Would a syllable with an f0 only two semitones higher than the expected mid level tone (or one semitone lower than the expected high level tone) still be heard as having a high level tone? In order to explore the relationship between the degree of frequency shift and the operation of tone normalization, three different degrees of pitch shift were used (one half semitone, one semitone, and two semitones).

A. Method

1. Participants

Twenty native Cantonese speakers (ten women and ten men) with no reported speech or hearing disability were recruited for the experiment. The mean age of the participants was 21.7 (range=20–24). Ten (five women and five men) were students in Speech and Hearing Sciences at the University of Hong Kong, with some training in phonetics. The remaining participants (five women and five men), were recruited from other faculties (schools) of the University of Hong Kong (Science, Arts, Law, Business, and Dentistry) and had no training in phonetics.

2. Stimuli

A Cantonese sentence: /ŋɔ23 wui23 tɔk22 ji33 pɛi25 lɛi23 t^hɛŋ55/我會讀意俾你聽 "I will read ji3 for you" modeled after the natural production of a native Cantonese male speaker (aged 22) was synthesized using Sensyn, a Klatt-style formant synthesizer from Sensimetrics Corporation (Klatt, 1980; Klatt and Klatt, 1990). To generate the synthetic stimulus, the natural sentence was sampled at 10 ms intervals and measurements of f0, amplitude envelope, and the first four formant frequencies were recorded and used as input to the Klatt synthesizer. A sentence was chosen that had the target syllable in the middle of the carrier phrase to avoid any potential for interaction with intonational effects on f0 related to the beginning or end of a phrase (cf. Vance, 1976). Six additional versions of the sentence were synthesized, with the nontarget portions of the sentence raised or lowered by $\frac{1}{2}$, 1, or 2 semitones while the target syllable itself remained unchanged (retained its original, natural f0 value of an average of 115 Hz). This resulted in a total of seven different carrier phrases, each surrounding a single syllable. These were designated as stimuli in the dynamic context condition. In addition, a monotone context version was created for each of the seven dynamic sentences by setting f0 to a uniform level equal to the average of the whole of each individual sentence (e.g., 116.0 Hz for the

TABLE I. F0 of each context sentence in the monotone context and average f0 of each context sentence in the dynamic context (experiment 1).

Direction of shift	Degree of shift	F0 (Hz)
Raised	2 semitones	131.1
	1 semitone	122.8
	$\frac{1}{2}$ emitone	119.4
Unshifted	0 semitones	116.0
Lowered	$\frac{1}{2}$ semitone	112.6
	1 semitone	109.2
	2 semitones	100.9

unshifted sentence). Thus, there were fourteen sentences in all: seven dynamic sentences (three with f0 raised by $\frac{1}{2}$, 1, or 2 semitones and three with lowered f0 (again, by $\frac{1}{2}$, 1, or 2 semitones) and one with an unshifted dynamic context), and seven monotone sentences corresponding to each of the dynamic sentences. Following Wong (1998) and Wong and Diehl (2003), in the raised conditions the target was expected to be identified as having a low level tone, while in the lowered conditions the target was expected to be identified as having a high level tone. Table I shows the f0 of each stimulus in the monotone context, calculated from the average f0 of the corresponding dynamic sentences.

3. Procedures

The experiment was carried out in a single-walled IAC sound-attenuating booth and took approximately 20 min to run. Stimuli were presented to listeners through Sennheiser HD-545 headphones, connected to an Apple PowerMacintosh 7100 computer. A Hypercard program was used for running the experiment. Before hearing the stimuli, an experimenter introduced and read aloud the three possible responses /ji55/醫 (doctor), /ji33/意 (*meaning*) and /ji22/ =(two) to the listeners, to ensure they were familiar with them.[4] A single block of fourteen stimuli was presented eleven times to each listener, for a total of 154 trials. Each block of trials contained the seven sentences from both the monotone and dynamic stimulus sets. All stimuli were presented in random order within each block.

In each trial listeners first heard a single sentence. Then the three possible responses were displayed on the screen above three numbered buttons. Listeners were asked to identify which of the three words appeared in the stimulus sentence by clicking one of the three buttons. The first block of trials (the first 14 trials) was treated as practice and results were not analyzed, although listeners did not know this at the time of the experiment.

B. Results and discussion

Shifting the fundamental frequency of a context sentence changed listeners' identification of a target syllable in the predicted manner: Downward shifts resulted in more high level responses, while upward shifts resulted in more low level responses. In order to more easily make comparisons across shift conditions, each response was scored according to whether it was the expected response for that condition (lowered, normal, and raised). For example, a response of /ji55/ (high level) was expected in the lowered context, while /ji22/ and /ji33/ were the expected responses in the raised and unshifted context conditions, respectively. The mean number of expected responses for each condition for each particular listener were then calculated and used for further analysis. No difference was found between participants with phonetics training and those without (mean =72% expected responses for both) or between male (mean=71%) and female participants (mean=72%). Therefore, these groups were combined for all further analyses.

Across all conditions, the degree to which responses matched predictions based on direction of shift was roughly proportional to the amount of shift (large, 2 semitone shift =100% expected response; medium, 1 semitone shift=74%; small, $\frac{1}{2}$ semitone shift=32%).[5] Two semitone shifts resulted in perfect or near perfect performance in the predicted direction, suggesting that listeners' expectations for tone locations in pitch space may be somewhat broader than might be predicted on the basis of Wong (1998) and Wong and Diehl's (2003) assumption, based on Chao's (1947) work, that a three semitone downward shift was necessary to induce listeners to hear the target as a high level tone. That is, while listeners may expect a high level tone to be approximately three semitones higher than a midlevel tone, they were equally willing to accept a tone that is only two semitones higher as a good high level percept. However, this is a tentative conclusion because of the presence of a ceiling effect, and the absence of some estimate of relative goodness.

Because of the ceiling effect in the two semitone pitch shift condition, the two semitone condition was excluded from further analysis. The number of expected responses for the unshifted condition was also quite high (not unexpectedly, mean=91% in monotone, and 94% in dynamic conditions). Therefore, this condition was also excluded from the analysis to simplify the data analysis. The proportion of expected responses for 1 and $\frac{1}{2}$ semitone shifts in both directions of both monotone and dynamic contexts is shown in Fig. 2.

A three-way within-subjects analysis of variance with factors context (monotone versus dynamic), direction (lowered versus raised), and size (one semitone versus one half semitone) was carried out. As expected, results showed a significant effect of size of shift, $F(1,19)=236.04$; $p<0.001$, such that there were more expected responses in the one semitone pitch shift condition (74%) compared to the half semitone shift condition (32%). There was also a significant effect of context, $F(1,19)=27.85$; $p<0.001$, such that the number of expected responses in the monotone context condition (mean=76%) was greater than in the dynamic context condition (mean=68%). This strongly suggests that listeners do not require prior experience with the complete range of a particular talker's pitch, but rather are able to compute the expected locations in pitch space of each lexical tone (at least the level ones) on the basis of the talker's average f0 alone. Similarly, there was a significant effect of direction, $F(1,19)=23.51$; $p<0.01$, such that there were fewer expected responses in the lowered condition (mean =59%) than the raised condition (77%). This was expected

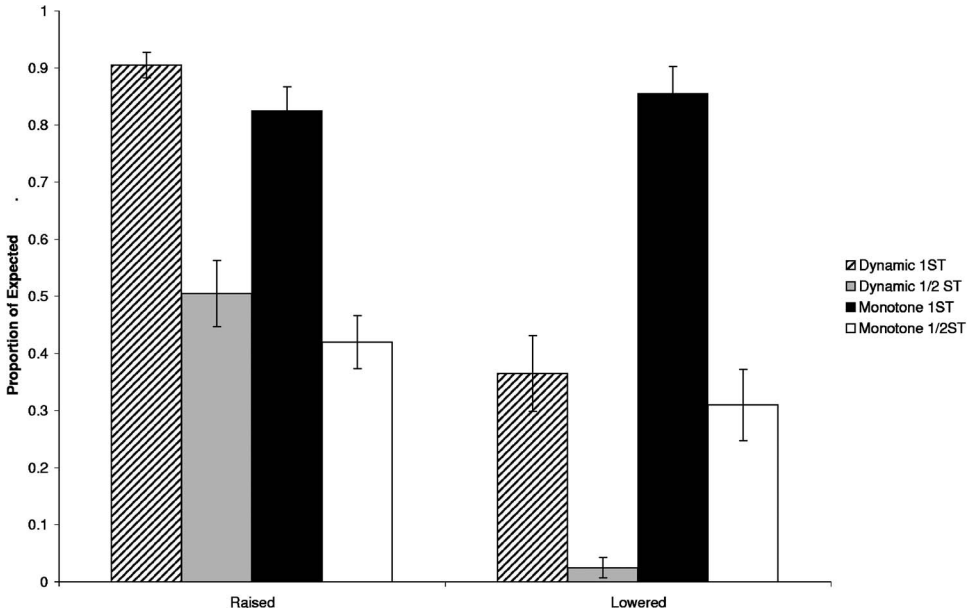


FIG. 2. Mean percentage of expected responses of 1 semitone and $\frac{1}{2}$ semitone shifts for monotone and dynamic contexts. Error bars indicate standard error.

because the two-semitone distance between a low level and a midlevel tone means that the maximum effect of raising would be reached within two semitones of upward shifting. In contrast, the *three-semitone* distance between midlevel and high level tones means that the maximum effect of lowering would only be reached with *three* semitones of (downward) shifting. Therefore, all else being equal, the raised condition should show more expected responses than the lowered condition, and this was indeed the case overall, especially for the dynamic context stimuli (71% expected response in the raised condition versus 20% in the lowered).

Unlike the dynamic condition, performance in the monotone condition was not affected by the direction of f0 shift (62% and 58% expected responses, respectively, and this was not significant by Tukey HSD post-hoc analysis, $p > 0.05$). The differential effect of monotone and dynamic f0 contours on the direction of F0 shift was supported by a significant interaction between context and direction, $F(1,19)=61.18$; $p<0.001$. Post hoc (Tukey HSD) analysis of this interaction revealed that all pairwise contrasts showed a significant difference ($p<0.05$) except for the one between the raised and lowered conditions in the monotone context, and the one between the dynamic and monotone contexts in the raised condition. This pattern of results suggests that something about the monotone context actually *improves* the overall effect of lowering the f0 of the context sentence. It is possible that the "robot-like" timbre of the synthetic speech might have encouraged listeners to treat the monotone (also "robot-like") context as somehow more acceptable, and therefore more effective, than the more dynamic context.

III. EXPERIMENT 2

The Pitch Range Assessment Model implies that optimal tone normalization can only be accomplished when the target syllable is preceded by a context containing speech at the upper and/or lower ends of the talker's pitch range. Experiment 1 of the present study demonstrated that it is not necessary to experience the full range of a talker's f0 range; exposure to the average f0 may suffice. Moreover, previous research suggests that tone normalization can exploit pitch information from either preceding (Wong, 1998; Wong and Diehl, 2003) or following (Lin and Wang, 1985) contexts, or both (the present experiment 1). However, it is still not clear how the temporal relationship between the context and the target affects tone normalization. Experiment 2 was designed to examine the relative importance of preceding versus following sentential context in lexical tone normalization.

A. Method

1. Subjects

Twelve college-aged, native Cantonese speakers (3 men and 9 women) with normal hearing and no history of speech or language disorder participated in this study.

2. Stimuli

The same semantically neutral sentence used in experiment 1 [/ŋɔ23 wui23 tɔk22 ji33 pɛi25 lɛi23 t^hɛŋ55/ "I will read *ji* for you (to hear)"] was recorded by a male native Cantonese volunteer in an IAC single walled sound-attenuating booth. Recordings were made via a Macintosh external microphone and recorded directly to disk using Sound Scope 16 (GW Instruments) via the built-in sound card of an Apple Macintosh G3 at a sampling rate of 44.1 kHz. Using Praat 3.9.27 (Boersma and Weenink, 2001), the stimulus was first low-pass filtered at 8k Hz. Subsequently, the f0 of the preceding and following contexts (those parts of the context sentence either preceding or following the word [ji]) was either (i) raised by one semitone, (ii) lowered by one semitone, or (iii) kept unshifted. Based on the results of experiment 1, a one-semitone shift was selected in order to avoid ceiling or floor effects in identification. All possible combinations of raising, lowering, and nonshifting were created, resulting in a total of nine conditions (3 levels of preceding context $\times 3$ levels of following context).

3. Procedure

The experiment was conducted in an IAC single-walled sound booth. The order of the nine stimuli in each block was randomized and 11 blocks were presented in all, resulting in a total of 99 trials for each participant. A Hypercard stack was used to present stimuli and collect responses. On each trial, a single stimulus was played out at 44.1 kHz via Sennheiser HD-545 headphones connected to an Apple PowerMacintosh 7100 computer, and the participant was asked to identify the target word by clicking on one of the characters shown on a computer screen. The possible characters were the same as in experiment 1. The first block served as a practice block and was excluded from analysis although the participants were not informed of this at the time. Each experiment session lasted for approximately 20 min.

B. Results

Participants' responses were recorded and scored with a "count-sum" scoring system: If the response indicated that the participant heard a low level tone then the response was given a score of −1; mid level tone responses were scored as 0; and high level tone responses were given a score of +1. The scores for each stimulus for each subject were summed. This resulted in a bounded range of scores from −10 to +10 since there were 10 trials for each stimulus. The more negative the score, the more the low level tone dominated the responses. Similarly, the more positive the score, the more the high level tone dominated the responses. If the score was close to 0, either the mid level tone dominated the responses or the three kinds of responses appeared roughly equally, or there were just high level and low level tone responses in almost equal measure. This method of scoring was chosen instead of the proportion of predicted responses used in experiment 1 because of the difficulty of determining *a priori* what the predicted response would be in cases where the preceding and following contexts were shifted in opposite directions. Figure 3 shows the scores for all the conditions.

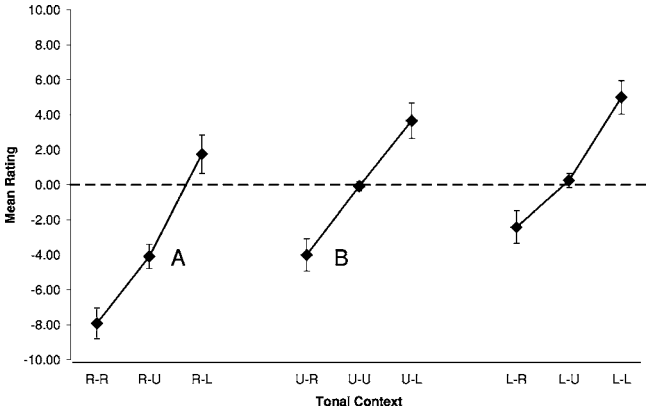


FIG. 3. Mean response scores (see the text) for tokens presented in contexts with varying patterns of shifted f0. R-R: raised precursor, raised following context; R-L: raised precursor, lowered following context; R-U: raised precursor, unshifted following context; L-R: lowered precursor, raised following context; L-L: lowered precursor, lowered following context; L-U: lowered precursor, unshifted following context; U-R: Unshifted precursor, raised following context; U-L: unshifted precursor, lowered following context; U-U: unshifted precursor, unshifted following context. More negative values indicate a higher proportion of low level responses, more positive values indicate a higher proportion of high level responses. Error bars indicate standard error.

As shown in Fig. 3, scores tended to increase when the context conditions changed from raised to unshifted to lowered, suggesting an increasing proportion of high level responses. In order to determine the relative effects of shifting the preceding versus following context, a two-way repeated measures ANOVA was calculated with three levels of preceding shift (raised, lowered, unshifted) versus three levels of following shift (raised, lowered, unshifted).

Results of the ANOVA showed that the effect of both the preceding and following contexts were statistically significant: Preceding, $F(2,22)=32.44$, $p<0.001$; Following, $F(2,22)=46.36$, $p<0.001$. On average, sentences with a raised preceding context were rated with a score of −3.4, as compared with 0.9 for lowered precursors and −0.1 for unshifted precursors. Similarly, sentences with a raised following context averaged −4.8, while lowered following contexts averaged 3.5, and unshifted following contexts averaged −1.3. The interaction between preceding and following contexts was not significant, $F(4,44)=2.08$, $p=0.10$.

Examination of means suggests that raising the following context had a stronger effect than raising the preceding context (raised following −4.8 versus raised preceding −3.4), while lowering the preceding and following contexts had similar effects (lowered following 3.5 versus lowered preceding 3.4). Because the experimental design means that cells with a raised precursor partially overlap with cells with a raised following context (e.g., the raised-raised sentences) these general patterns cannot be tested directly as part of the main ANOVA. However, it is possible to contrast the effects of raising and lowering the precursor and following contexts by comparing individual cells in the complete interaction. For example, when both the precursor and following context are raised, the mean response is most negative (−7.92), suggesting that listeners would most likely perceive the target word as having a low level tone in this condition. When neither context is shifted the mean score is virtually zero (−0.08), suggesting that listeners would most likely perceive the target word as having a midlevel tone in this condition. Finally, when both contexts are lowered, the mean response is the most positive (5.00), suggesting that listeners would most likely perceive the target word as having a high level tone in this condition. Raising the context increases the proportion of low level responses, while lowering it increases the proportion of high level responses, and the overall magnitude of the raising effect is somewhat larger than that of the lowering. Thus, these results correspond quite well to the findings of the previous experiment.

By examining specific pairs of context combinations using post-hoc (Tukey HSD, $\alpha=0.05$) analysis it is possible to tease apart the separate effects of raising and lowering the f0 of both preceding and following contexts. For example, when the precursor was raised and the following context was unshifted, the mean score was −4.08, but when the following context was raised and the precursor was unshifted the mean score was −4.00 (comparison of points *A* and *B* in Fig. 2), and this difference was not significant. In contrast, there

were significant differences between the effects of lowering the precursor as compared with the following context (0.25 precursor lowered versus 3.66 following context lowered; compared to the unshifted precursor and unshifted following context score of −0.08). These results suggest that lowering the precursor alone had little or no effect on listeners' responses (no different than the unshifted/unshifted context). In contrast, raising either the precursor or following context had a strong effect, as did lowering the following context.

In order to determine whether the precursor or following context had the stronger effect, the raised-lowered (RL) and the lowered-raised (LR) conditions were compared. If the precursor and following contexts had the same degree of effect on tone perception, then the responses to these two conditions should be identical. However, as seen in Fig. 3, responses to the RL condition were generally slightly positive (mean=1.75), indicating a greater prevalence of high level responses (consistent with the following context), while responses to the LR condition were somewhat more negative (mean=−2.42), indicating more low level responses (again, consistent with the following context). The difference between these two conditions was significant according to post-hoc (Tukey HSD) analysis.

A similar comparison can also be carried out by computing the magnitude of the effect of shifting only one of the two contexts while leaving the other context unshifted. For example, the difference between the raised following context (with unshifted precursor context, UR) and the lowered following context (with unshifted precursor context, UL) gives an effective magnitude of 7.67 for shifting the following context. In contrast, the magnitude of the effect of shifting the preceding context when the following context is held constant at the unshifted value was 4.33 (RU versus LU). Although there is not a significant difference between these two magnitudes (7.67 vs 4.33), $t(11)=1.98$, $p=0.07$, the overall pattern supports the hypothesis that the following context had a greater influence than the preceding context on normalization of Cantonese lexical tones.

C. Discussion

Results of experiment 2 suggested that both raising and lowering both the preceding and the following contexts had significant effects on Cantonese tone normalization. As in experiment 1, the effect of raising the f0 of the context one semitone was greater than that of lowering it by the same amount. Furthermore, the effect of changing the following context was larger than that of changing the preceding context. When the two contexts gave contradictory information, listeners appeared to depend more on the following context to perform tone normalization.

This pattern of results is broadly consistent with the findings of Lin and Wang (1985), in that they also found an effect of following context on the perception of lexical tone. In addition, other kinds of normalization, including phonetic context normalization (Johnson and Strange, 1982) and speaking rate normalization (Hirata and Lambacher, 2004; Newman and Sawusch, 1996) have been shown to be affected by subsequent as well as preceding extrinsic context.

Following these results, the Pitch Range Assessment Model may need to be extended to include some reference to the role of context location (in addition to the changes indicated by the results of experiment 1). Specifically, the results of experiment 2 suggest that listeners give more weight to the following context than to the preceding context, in a manner similar to other aspects of auditory pitch processing (Brady, *et al.*, 1961; Ciocca and Darwin, 1999). One might conceptualize such a process either as a moving window of analysis (e.g., a running average) or in terms of a feedback loop system. In a moving window type of system, the listener tracks the average f0 within a particular time frame (the window width), but the window moves through the signal over time. As the window moves away from earlier-occurring portions of the signal their importance declines, while the relative contribution of later-occurring information increases as the window moves toward/over them. Obviously, for post-target information to play a role in determining the target syllable's tone the later-occurring information must have already occurred by the time the decision is made. Thus, implicit in this model is the idea of waiting until some (unspecified) time after the target to make a final decision regarding the identity of that target [cf. discussion of speaking rate normalization by Miller and Dexter, (1988)].

Such a process is made more explicit in a feedback-loop based system. In a feedback-based system (cf. Nusbaum and Schwab, 1986) the listener is presumed to be continuously updating a hypothesis about the identity of the target on the basis of both internally generated and externally available information. Thus, the hypothesized identity of the target syllable could change as a consequence of later-occurring information.[6] On the basis of the present evidence, it is not possible to distinguish between the two models. Furthermore, it is not entirely certain that either model is necessarily involved in natural speech processing. Indeed, it is possible that listeners' preference for using later-occurring information to judge the identity of a medially presented target syllable may derive from the nature of the experimental task: Hear a sentence and report the identity of a given syllable. When carrying out this task it seems highly likely that listeners might (consciously or unconsciously) choose to reserve judgment until the entire sentence has been presented. Whether or not this kind of delayed processing takes place in more natural speech perception situations remains to be determined.

Finally, it is also possible that the preference for using later-occurring information in the present experiment provides support for the original Pitch Range Assessment Model, albeit in a more complex form. In the present experiment the following context had a larger pitch range than the preceding context. Tone number values ranged from 2 (low) to 5 (high) in the following context, but only from 2 to 3 (mid) in the preceding context. Thus, the following context provides a more complete picture of the pitch range of the speaker than does the preceding context. Thus, it is possible that listeners chose to reserve judgment about the identity of the target syllable because they knew that the following context would provide more information about the overall pitch range of the talker. Future experiments might control for this

possibility by providing equivalent tonal information in both preceding and following contexts. Still, all possible interpretations of these results suggest that, when possible, later-occurring information is incorporated into the tonal decision-making process, and is even preferred under certain task and stimulus conditions.

IV. EXPERIMENT 3

Research on phonetic context effects in speech perception frequently addresses the question of whether contextual effects derive from mechanisms specific to the perception of speech, or whether they result from more general auditory processes of assimilation and/or contrast that operate independently of whether the sounds being perceived are human speech or not (cf. Holt *et al.*, 2000). Similarly, it is possible that the talker-dependent normalization of lexical tones might result from the operation of a more general auditory process related to the perception of the pitch of complex tones or voices more generally, rather than one specifically related to the perception of linguistic tones. Wong (1998) showed that tone normalization could occur even across two languages, but in that case the listeners were familiar with both of the languages they heard, and both the context and target constituted meaningful utterances in one of the two languages. In order to determine whether normalization of lexical tones depends on listeners' ability to process the speech signal in a linguistically meaningful manner, the third experiment evaluated normalization of lexical tones presented within a context that could not be interpreted linguistically.

A. Methods

1. Subjects

Twelve college-aged, native Cantonese speakers (3 men and 9 women) with normal hearing and no history of speech or language disorder participated in this study. Note that these participants were recruited simultaneously with those in experiment 2, and assignment to either experiment 2 or experiment 3 was done on a gender-matched (but otherwise random) basis. None of these participants had taken part in any previous study on lexical tone normalization.

2. Stimuli

Stimuli were identical to those used in experiment 2, except that the context portions of each stimulus were rendered unintelligible by extracting the f0 contour and resynthesizing it using the "hummed" neutral vocal tract ([ə]) function in Praat 3.9.27.[7] As in experiment 2, there was a total of nine context conditions (3 levels of preceding context $\times$3 levels of following context).

3. Procedure

Experimental procedures were identical to those used in experiment 2.

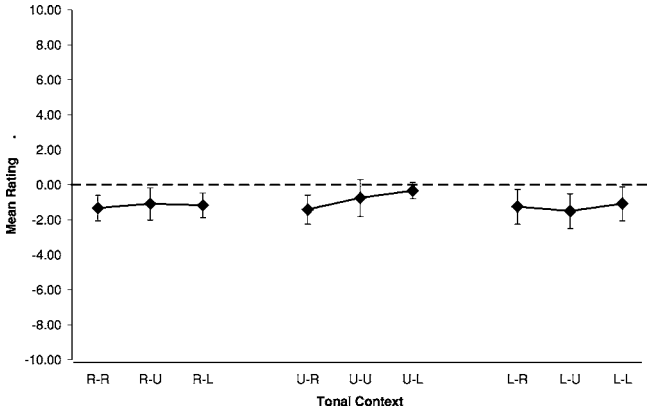


FIG. 4. Mean response scores (see the text) for tokens presented in unintelligible (/ə/) contexts with varying patterns of shifted f0. R-R: raised precursor, raised following context; R-L: raised precursor, lowered following context; R-U: raised precursor, unshifted following context; L-R: lowered precursor, raised following context; L-L: lowered precursor, lowered following context; L-U: lowered precursor, unshifted following context; U-R: unshifted precursor, raised following context; U-L: unshifted precursor, lowered following context; U-U: unshifted precursor, unshifted following context. More negative values indicate a higher proportion of low level responses, more positive values indicate a higher proportion of high level responses. Error bars indicate standard error.

B. Results

Responses were scored as in experiment 2, and results are shown in Fig. 4. As in experiment 2, a two-way repeated measures ANOVA was calculated with both direction of shift (raised, unshifted, lowered) and location of shift (preceding, following) as factors. In this case, results showed no significant effects of any kind. Listeners gave far more mid level responses (53.6%) than either low level (28.7%) or high level (17.7%) responses, suggesting that they were responding primarily to the absolute fundamental frequency of the target syllable (originally produced as a mid level tone) independently of the f0 of the surrounding context.

C. Discussion

In this experiment, listeners seemed unable or unwilling to use linguistically meaningless context for the purposes of tone normalization. It is possible that listeners treat nonword stimuli differently for the purposes of tone normalization, but previous studies in other languages have showed that listeners are able to make linguistically informed judgments about pitch patterns even when producing or perceiving nonsense words (e.g., Pierrehumbert, 1979). Thus, it may be assumed that listeners in the present study were able to perceive the pitch patterns of the context sentence, but for some reason failed to use the schwa-only context as a cue to the talker's fundamental frequency for the purposes of tone normalization. It is less clear whether listeners were consciously or unconsciously tuning out the schwa context (if such conscious control over a seemingly basic speech perceptual phenomenon is even possible). However, in either case, it seems plausible that some property of the context failed to match sufficiently with the corresponding property in the target syllable, allowing listeners to dissociate (consciously or unconsciously) the two types of signals and consider only the target syllable in making their tonal judgment. There are at least

three possible ways in which the context and target were mismatched, one linguistic and one nonlinguistic.

First, it is possible that listeners did not hear the [ə]-resynthesized stimuli as speech at all. While previous research suggests that nonspeech sounds are capable of supporting other kinds of (segmental) contextual normalization processes (Holt *et al.*, 2000), it is not clear that tone normalization necessarily operates according to identical principles. Further research is needed to determine whether or not manifestly nonspeech signals (e.g., pure tones) can function as a context for the purposes of tone normalization. Second, even if listeners did hear the hummed context as speech, the sound [ə] is not a phoneme in Cantonese, and therefore listeners may have ignored it because it is too "foreign" to be relevant for making tonal decisions. This is possible, but not likely, as Wong (1998) has already shown that a familiar foreign language context (English) can provide sufficient pitch information to support at least some degree of lexical tone normalization. Finally, it is possible that listeners treated the schwa-context as if it had been produced by a different talker. In the present stimuli, no attempt was made to match any talker identity-related parameters of the context to that of the target. Thus, it is possible that lexical tone normalization depends on the perception that both the context and the target were produced by the same talker.

V. EXPERIMENT 4

The process of talker normalization represents an interface between processes of talker identification and recognition on the one hand, and processes of speech perception on the other. However, there has been relatively little research on the degree to which the perception of talker *identity* influences talker normalization.[8] In some studies of tone normalization, all tokens were produced by the same talker (e.g., Wong, 1998; Wong and Diehl, 2003). In those cases in which talker and target were not produced by the same talker [e.g., Moore and Jongman (1997); Leather (1983)], synthetic stimuli were presented within a sentential context produced by a natural talker. However, both of these studies attempted to synthesize the target tokens such that there was little possibility that listeners might identify the target as having been produced by a different talker than the context. If tone normalization depends on the operation of mechanisms for talker identification, then we would expect a perceptible difference in talker between context and target to disrupt it. If hearing a target and context produced by a different talker disrupts the process of tone normalization even when the context consists of linguistically meaningful, natural speech in the listeners' native language, this would provide strong support for the hypothesis that tone normalization (and, by extension perhaps talker normalization in general) are closely related to processes of talker recognition or identification. On the other hand, if tone normalization occurs across obviously different talkers, this might suggest that perceptual normalization, at least for tone, may rely on the operation of more general auditory mechanisms that are not specific to the perception of speech *per se*.

A. Methods

1. Subjects

Twenty-two native Cantonese speakers (16 women, 6 men, aged 21–24 years) with normal hearing volunteered to participate in this experiment. All participants were students in the Division of Speech and Hearing Sciences at the University of Hong Kong.

2. Stimuli

Two native Cantonese men who were judged by P.C.Y.C. (a final-year student in speech and language pathology who had received phonetic training) as having similar pitch ranges, but clearly different-sounding voices, were selected as speakers. They read the same semantically neutral context sentence with embedded mid level tone and stimuli were recorded using the same procedures as in experiment 2. The average f0 for the carrier phrase with the mid level tone was 123.9 Hz for speaker 1 and 112.3 Hz for speaker 2. One production of the target sentence with the target tone for each speaker was selected by two final-year speech and language pathology students (one of whom was P.C.Y.C.). The two sentences were selected so that the target word could clearly be identified as a mid level tone, and both sentences had a perceptually similar rhythm. The three segments of the carrier sentence (preceding context, target syllable, and following context) were measured for both speakers to ensure that they had similar durations across the two speakers: the duration of the preceding context was about 710 ms for speaker 1 and 695 ms for speaker 2; the following context was about 720 ms for speaker 1 and 745 ms for speaker 2. The duration of the target word was set to 250 ms (close to the duration of the natural token produced by speaker 2) for both speakers using the manipulation functions of the Praat software. Measurements of f0 for each word for each sentence were made, and speaker 1 was identified as having a larger overall range of f0 values across the sentence, and was designated as the model talker.

After this, the amplitude of each of the two sentences (one per speaker) was peak-normalized, and the fundamental frequency pattern of each word within each sentence was resynthesized using the PSOLA algorithm of the Praat software. For resynthesis, the f0 values of the two sentences were first extracted using the default autocorrelation algorithm in Praat. Speaker 2's f0 values for each word within each sentence were then manually adjusted to new f0 values that were close to the f0 values for that word of speaker 1 (the model talker). Note that, although the average f0 of the resynthesized target (106 Hz) was more similar to the average context f0 of speaker 2, this was because the range of speaker 1 was larger. Therefore, the words /teng55/ and /bei25/ ended at a higher f0 for speaker 1 than speaker 2, which is why the average f0 of the sentence context was higher for speaker 1. Exact F0 values were set such that the sentences retained a natural-sounding intonation and the target word had correct tonal identity as judged by P.C.Y.C. and two other native Cantonese speakers. The carriers for both speakers were then resynthesized with the new, similar F0 values using the PSOLA algorithm. After resynthesis, the F0

TABLE II. Fundamental frequencies of stimuli after modification (experiment 3). Note: Entries in italic indicate stimuli presented to the same talker group. Entries in Roman indicate stimuli presented to the different talker group.

	Fundamental frequency (Hz)		
Context	Carrier phrase	Target word talker 1	Target word talker 2
Talker 1: raised	97–133	*106*	106
Talker 1: unshifted	92–125	*106*	106
Talker 1: lowered	86–118	*106*	106
Talker 2: raised	96–133	106	*106*
Talker 2: unshifted	91–125	106	*106*
Talker 2: lowered	86–117	106	*106*

values of the sentences for the two speakers were determined to be very closely matched (within a few Hertz for each word): The average f0 of the target words in the resynthesized stimuli was 106 Hz for both speakers.

After generating these sentences of matched duration and f0 characteristics, two additional sentences were created by splicing the target word from speaker 1 into the carrier phrase spoken by speaker 2, and vice versa, thereby creating two additional stimulus sentences in which the target syllable was produced by a different talker than was the context in which it appeared. Thus, there were four base sentences differing only in terms of the (nonpitch) vocal properties: Two with matching talkers for context and target, and two with mismatched talkers for context and target. The f0 values of the context (nontarget) portion of each of these four base sentences were then either raised, unshifted or lowered by one semitone as in experiment 1. F0 properties for the resulting twelve stimuli are shown in Table II.

3. Procedure

Participants were randomly assigned to one of two groups. In the same talker group, listeners heard only sentences in which the same talker produced the target syllable and the context in which it was embedded. Listeners in the different talker group heard only stimuli in which the target syllable and surrounding context were produced by different talkers. Experimental procedures were otherwise identical to those of Experiment 2.

B. Results

As in the first experiment, the proportion of expected responses was calculated for each condition. A three-way, mixed factorial ANOVA was calculated, with the type of carrier phrase (Same Talker versus Different Talker) as a between groups factor and the direction of f0 shift (raised, lowered, unshifted) and target talker (talker 1, talker 2) as within groups factors. Results of the ANOVA indicate no significant main effects of type, $F(1,20)=1.18$, $p=0.29$, or talker, $F(1,20)=0.88$, $p=0.36$. There was a significant effect of shift, $F(2,40)=7.20$, $p=0.002$, such that lowered contexts resulted in 53% expected (high level) responses, unshifted contexts result in 80% expected (mid level) responses, and raised contexts result in 70% expected (low level) responses.

Significant two-way interactions were found between talker and type, $F(1,20)=11.43$, $p=0.003$, shift and type, $F(2,40)=5.17$, $p=0.01$, and talker and shift, $F(2,40)=10.15$, $p<0.001$.

These interactions can be explained by exploring the significant three-way interaction between talker, shift, and type, $F(2,40)=6.55$, $p=0.003$, as shown in Figs. 5(a) and 5(b). Post-hoc (Tukey HSD, $\alpha=0.05$) analysis revealed that the primary sources of the observed interactions were (1) a significantly lower proportion of expected responses in the same talker, lowered context conditions (leftmost circles in each graph, and significantly lower than either the raised or unshifted same talker conditions), and (2) a significantly lower proportion of expected responses in the raised, different talker condition with the target produced by talker 2 (rightmost closed squares in Fig. 5(a), and significantly lower than for the corresponding condition of tokens produced by talker 1 but presented within a talker 2 context).

VI. GENERAL DISCUSSION

The results of the same talker condition (T1/T1 and T2/T2) in the present experiment were comparable to those found in experiments 1 and 2, such that shifting the f0 of the context upward resulted in more low level responses, while shifting it downward resulted in more high level responses and the effect of the downward shift was not as strong as the upward shift. However, the pattern of responses in the different talker condition provides additional insight into the process of tone normalization. First, we observed that, in the different talker condition (T1/T2 and T2/T1), listeners were overall more inclined to respond as expected, and this appears to be primarily the result of an increase in the effectiveness of the lowered shift condition in the different group as compared with the same talker group: Both the T1/T2 and T2/T1 stimuli showed a greater proportion of expected responses with a lowered context than did either the T1/T1 or T2/T2 conditions in the same context. Second, although post hoc (Tukey HSD, $\alpha p=0.05$) tests revealed no difference between shift conditions when averaging across talkers, in the raised condition there is evidence of an effect of individual talkers. When the talkers who produced the target and context did not match (squares in both graphs), in the raised condition there was a significant drop in expected responses to talker 2 targets in talker 1 context as compared with talker 1 targets in talker 2 context. This was the only significantly different comparison between the two talkers. Thus, mismatching the talker of the target and the context seems to improve listeners' normalization of lexical tones in the lowered context condition, but in the raised condition normalization is disrupted when talker 1 targets are placed in a talker 2 context (but not vice versa).

The disruption of normalization when talker 1 targets are presented in a talker 2 context could be explained by differences in the sound of the speech of the two speakers. Although the f0 of the two speakers in the stimuli was adjusted to be identical, the rest of the speech [the perceived

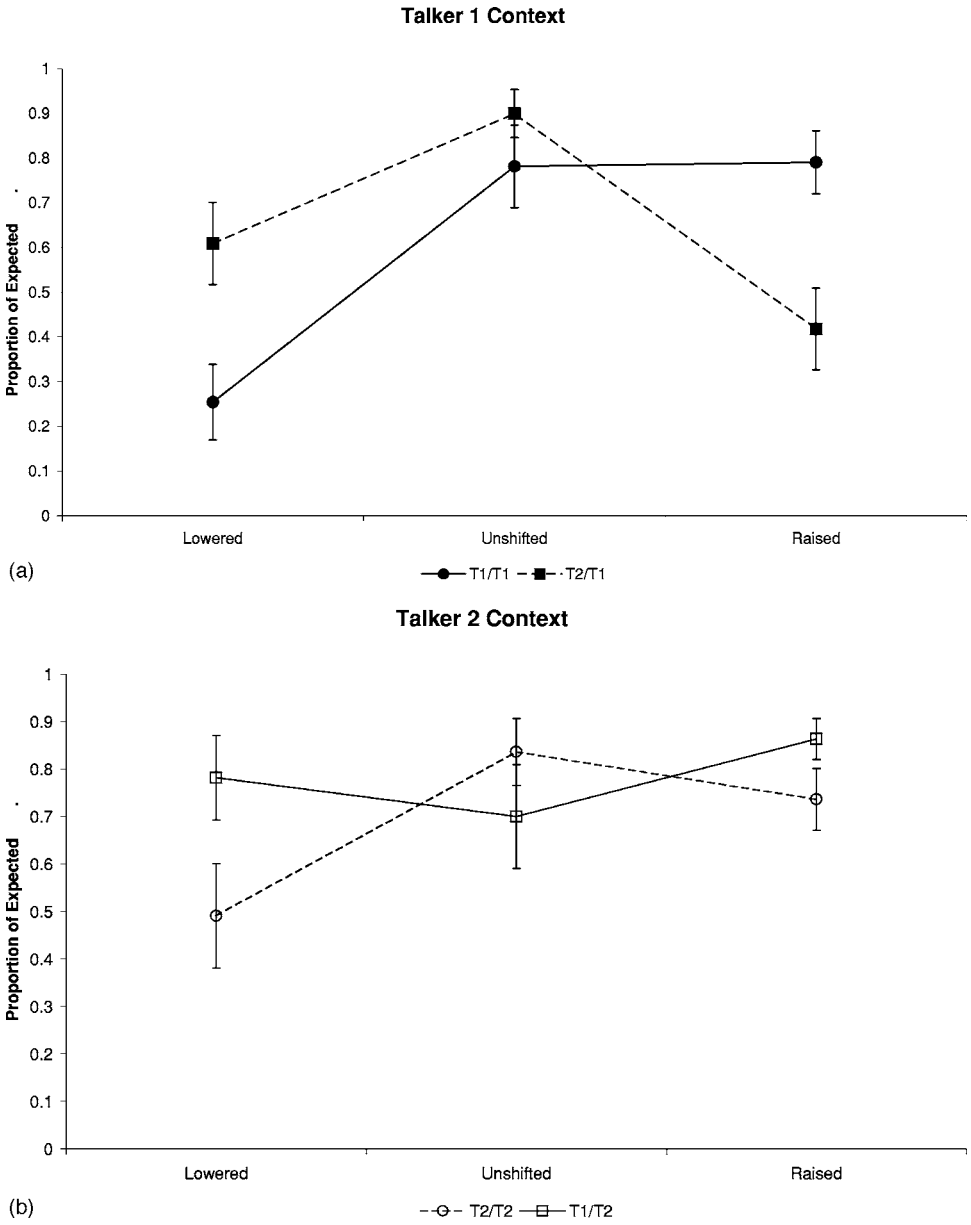


FIG. 5. (a) Mean proportion of expected responses for targets produced by talker 1 (T1/T1, circles, solid line) or talker 2 (T2/T1, squares, solid line) embedded within context sentences produced by talker 1. Labels first list the talker who produced the target, followed by the talker who produced the context. In (a) and (b), stimuli with targets produced by talker 1 are indicated with solid lines, while stimuli with targets produced by talker 2 are indicated with dashed lines. Stimuli with contexts produced by talker 1 are indicated with closed symbols, while contexts produced by talker 2 are indicated by open symbols. Stimuli where the same talker produced both target and context are indicated with circles, while those in which the talker of the target and context differ are indicated with squares. Error bars indicate standard error. (b) Mean proportion of expected responses for targets produced by talker 1 (T1/T2) or talker 2 (T2/T2) embedded within sentences produced by talker 2. Nomenclature and symbols as in (a). Error bars indicate standard error.

"timbre," including properties related to both laryngeal (voice quality, source spectrum) and supralaryngeal differences (vowel formant spacing, relative timing of articulators)] of each talker was (intentionally) noticeably different. Singh and Hirsh (1992) found that variation in timbre could be perceived as a change in pitch when stronger pitch cues such as the fundamental frequency were kept constant. In the present experiment, ten additional listeners exposed to the same talker stimuli rated talker 2 as having a higher pitched voice than talker 1 even though the f0 of the two talkers' stimuli was identical. Since the timbre of a speech signal depends on both the formant (filter) properties and the harmonic (source) properties of the talker's voice, and the source properties were identical, this suggests that differences in the two talkers' individual formant patterns must contribute to the difference in their perceived pitch. The source/filter resynthesis algorithm used for shifting f0 in the present experiment does not alter formant frequencies. Thus, each talker's filter function remained constant across shift conditions. That is, talker 1's speech must have been perceived as having an overall lower pitch than talker 2's speech in all conditions. Therefore, when the target syllable produced by talker 2 was placed within the raised context produced by talker 1, the perceptual magnitude of the f0 shift effect was reduced because of the higher pitch perceived for talker 2 as compared with talker 1.

This effect of timbre, however, cannot explain the overall superior effect of the lowered shift in the different talker condition as compared to the lowered shift in the same talker condition. Indeed, post-hoc (Tukey HSD, $\alpha=0.05$) analysis showed that listeners in the lowered shift condition who heard talker 1 targets presented in talker 2 context showed a greater proportion of expected (high level) responses ($p=0.05$) compared with those who heard talker 1 targets in talker 1 context, despite the fact that timbre differences should have caused a smaller perceived effect of shift in the T1/T2 condition because the talker 2 context should have sounded less lowered than the corresponding talker 1 context

in the same (lowered) shift condition. Conversely, we would have expected timbre differences to produce an enhancement of the shift effect in the raised condition when presenting talker 2 targets in talker 1 context (T2/T1) as compared to the T2/T2 presentation because the perceived pitch of the talker 2 target should have been lower, or the perceived pitch of the T1 context should have been higher, than the f0 measurements alone would have indicated, or both. However, there was no significant difference in proportion of expected responses to talker 2 targets in talker 1 versus talker 2 contexts.

The specific details of how timbre affects the perception of pitch for the purposes of normalization of tones remains to be explained, but the results of the lowered condition clearly show that normalization can actually be stronger when target and context do *not* match (as compared to when they do), supporting the hypothesis that tone normalization does not *require* the perceived continuity of talker.

If perceptual normalization of lexical tone does not require a perceived continuity of talker, why, then, did listeners fail to use pitch information from extrinsic context in experiment 3, where the words in the context had been replaced by [ə]? At this point, it seems reasonable to conclude that listeners were able to recognize the series of schwa-syllables as being meaningless, and were therefore able to treat them as irrelevant to tonal processing. This suggests that the process of tone normalization is, in some sense, also a linguistic one, and does not depend solely on the automatic operation of nonlinguistic, auditory processes. While the determination of pitch for the purposes of talker normalization appears to function independently of perceived talker continuity (and therefore may be the result of a more general auditory process), the application of such information to tone normalization may require that listeners recognize the context as being linguistically meaningful. However, these conclusions are only tentative, pending the results of further research on the ability of listeners to normalize lexical tones in the context of a talker speaking a real language, but one that is unknown to the listener [though see Jongman and Moore (2000) for preliminary results in this direction]. If future research demonstrates that even appropriately constructed nonspeech contexts can induce tone normalization effects, this would instead suggest that the phenomenon of tone normalization, and perhaps talker normalization in general, might derive from more basic auditory processes for maintaining perceptual constancy across contexts (see Holt, 2005 for similar discussion).

The application of general processes of perceptual constancy to the specific task of talker normalization might be universally successful under normal (real world) task conditions, and yet could still perform less than optimally in specific (typically laboratory-related) contexts. For example, in the real world it is relatively improbable that talker identity would change fluently in the middle of a sentence, and even if it did there would necessarily be some additional cues to the change including changes in the spatial location of the voice, as well as the strong possibility of additional differences in intrinsic vocal properties such as timbre and fundamental frequency. However, in the laboratory it is possible to artificially optimize the transition between context and target in order to preserve most cues to continuity (e.g., by matching fundamental frequencies and maintaining the same spatial location for the voice, as in the present experiment). Under such artificial circumstances listeners may be "fooled" into applying a valid, general perceptual strategy across talkers in the same way that many visual illusions result from the correct operation of visual strategies that are adequate for performance in the real world but yield predictably illusory percepts when "tricked" in particular ways in the laboratory (Marr, 1982, p. 30).

VII. SUMMARY

In order to accurately identify the tonal category of a given syllable, tone language speakers use pitch information obtained from extrinsic speech to estimate the expected pitch properties of possible tones (the talker's tone space). The perceived lexical tone of a syllable is based on a match between the perceived pitch of the syllable and the representation of that tone within the mental tone space induced for that talker on the basis of pitch information available from context. The results of the experiments presented here provide further information relevant to evaluating and modifying the Pitch Range Assessment Model. The original formulation of the PRAM proposed that listeners' estimate of a given talker's pitch space depends on the pitch range present in the extrinsic context. Results from the present experiment 1 suggest instead that the mechanism for estimating a given talker's tonal space involves a process of extrapolation from the talker's average f0. Previous studies of tone normalization examined the effects of either preceding context (e.g., Fox and Qi, 1990; Leather, 1983; Moore and Jongman, 1997; Wong, 1998; Wong and Diehl, 2003) or of a single post-target syllable (Lin and Wang, 1985) but did not specifically compare the two types of conditions to determine which (if either) had a greater effect on normalization. Furthermore, no previous study has specifically attempted to determine whether tone normalization is affected by perceived differences in talker. Results of the present experiments 2 and 4 suggest that pitch information for tone normalization appears to be integrated both across different regions of context (both preceding and following the target) as well as possibly across different sources of pitch perception (f0 and timbre), although the evidence supporting a role for timbre-based pitch information is only suggestive at this point. Finally, the results of the present experiments 3 and 4 suggest that the process of tone space estimation does not function in a talker-specific manner, but may still depend on the listeners' linguistic knowledge. That is, tone normalization does not seem to be dependent upon listeners' perception that the entire utterance was produced by a single talker, but it does appear to be necessary that the listener be able to recognize the context as meaningful speech. Whether or not the speech must be meaningful to listeners is still undetermined [though see Jongman and Moore (2000) for evidence that English listeners do show some perceptual normalization of Mandarin tones, albeit not the same pattern as Mandarin listeners].

Further research on the possible effects of unfamiliar linguistic contexts will be necessary to investigate this hypothesis more thoroughly.

ACKNOWLEDGMENTS

We would like to thank Yukari Hirata, Phil Rose, and two anonymous reviewers for helpful comments given on an earlier draft of this manuscript. The work described in this paper was funded by a grant from the Research Grants Council of the Hong Kong SAR, China (Project No. HKU 7913/00H).

[1]In order to compare tones across languages, tones in this paper are transcribed in a manner similar to the Chao tone number system used by Bauer and Benedict (1997). This system reflects the relative pitch of the syllable at onset and offset within a 5-point scale from the bottom (1) to top (5) of the talkers' normal pitch range. Thus, for the Cantonese tonal system 55 =high level, 25=high rising, 33=mid level, 21=low falling, 23=low rising, and 22=low level. For Mandarin, 55=tone 1 (high level), 35=tone 2 (rising), 214=tone 3 (dipping), and 51=tone 4 (Falling).

[2]Moore and Jongman (1997) actually used a continuum of syllables ranging in both frequency and time of the turning point of f0 from a good 25 (tone 2) to 214 (tone 3) syllable. The description of the effect they identified is given here with respect to a single syllable in the middle of the continuum in order to emphasize the role of contextual differences.

[3]Note that Wong (1998) was able to increase the effectiveness of the English context until it matched that of the Cantonese context by increasing the degree of difference between the average f0 of the context and that of the target. Cantonese listeners required a four semitone upward shift or a three semitone downward shift of the English context before they reported that the target had a low level or high level tone (respectively) to the same degree that they did with a three semitone upward or two semitone downward shift in a Cantonese precursor. These results not only suggest that listeners' response patterns are gradient (as confirmed in the present experiment 2), but also further support the hypothesis that listeners perform tone normalization differently depending on the language of the context.

[4]With respect to translation, note that the characters /ji33/意(meaning) and /ji22/ (two) are commonly produced as single words and thus translating them as such poses no problem. In contrast, the status of /ji55/醫is not as clear. This character does not typically appear outside of multisyllabic collocations, but its most common use is in the two-word combination that means *doctor* (followed by生/sa:ŋ55/). Thus, the closest English interpretation of the semantic contribution of /ji55/醫to this collocation is *doctor*. With respect to asking listeners to respond to single characters, native speakers have no trouble pronouncing the character醫/ji55/ in isolation. Moreover, in the present study listeners were instructed to respond on the basis of the *character* they heard. Thus, although the character 醫/ji55/ does not typically appear in isolation, the task of identifying this item was not markedly more difficult or unusual than identifying either of the other two tokens.

[5]Nonexpected responses in the raised and lowered conditions were generally 33 (mid level) responses, reflecting either (or both) the fact that the f0 of the target syllable was in fact originally that of a mid level tone, and/or that it is very unlikely to misidentify a high level tone as a low level tone, or vice versa. Thus, a mid level response was the most likely default response both on the basis of its intrinsic f0 properties, and simply because the midlevel tone lies between the high and low level tones.

[6]This process is also compatible with the theory of Analysis by Synthesis (e.g., Stevens and Halle, 1964).

[7]The Praat resynthesis menu selection reads "To Sound (hum)" but the actual result is much closer to a schwa than a hum. The resulting sound is clearly non-nasal, and has formant frequencies of F1=600, F2=1400, F3 =2392, and F4=3412, quite similar to those measured by Peterson and Barney (1952) for the central unrounded vowel of American English.

[8]Although see work by Nygaard and colleagues (Nygaard and Pisoni, 1998; Nygaard *et al.*, 1994) for important examples of work in this direction. This work does not, however, deal with the perception of lexical tones.

Bauer, R. S., and Benedict, P. K. (**1997**). *Modern Cantonese Phonology* (Mouton de Gruyter, Berlin).

Boersma, P., and Weenink, D. (2001). "Praat: A system for doing phonetics by computer." Retrieved May 22, 2001, from University of Amsterdam, Institute of Phonetics Sciences Web site: http://www.praat.org (confirmed 15 April, 2005).

Brady, P. T., House, A. S., and Stevens, K. N. (**1961**). "Perception of sounds characterized by a rapidly changing resonant frequency," J. Acoust. Soc. Am. **33**, 1357–1362.

Chao, Y. R. (**1947**). *Cantonese Primer* (Harvard University Press, Cambridge, MA).

Ciocca, V., and Darwin, C. J. (**1999**). "The integration of nonsimultaneous frequency components into a single virtual pitch," J. Acoust. Soc. Am. **105**, 2421–2430.

Fok Chan, Y. Y. (1974). "A perceptual study of tones in Cantonese," Centre of Asian Studies, University of Hong Kong, Hong Kong.

Fox, R., and Qi, Y. (**1990**). "Contextual effects in the perception of lexical tone," J. Chinese Linguistics **18**, 261–283.

Francis, A. L., Ciocca, V. C., and Ng, B.K. C., **2003**. On the (non)categorical perception of lexical tones," Perception & Psychophysics, **65**(6), 1029–1044.

Fu, Q.-J., Zeng, F.-G., Shannon, R. V., and Soli, S. D. (**1998**). "Importance of tonal envelope cues in Chinese speech recognition," J. Acoust. Soc. Am. **104**, 505–510.

Hirata, Y., and Lambacher, S. G. (**2004**). "Role of word-external contexts in native speakers' identification of vowel length in Japanese," Phonetica **61**, 177–200.

Holt, L. L. (**2005**). "Temporally non-adjacent non-linguistic sounds affect speech categorization," Psychol. Sci. **16**, 305–312.

Holt, L. L., Lotto, A. J., and Kluender, K. R. (**2000**). "Neighboring spectral content influences vowel identification," J. Acoust. Soc. Am. **108**, 710–722.

Johnson, K. (**1991**). "Differential effects of speaker and vowel variability on fricative perception," Lang Speech **34**, 265–279.

Johnson, K. (**1997**). "Speech perception without speaker normalization: An exemplar model," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix (Academic, San Diego), pp. 145–165.

Johnson, K. (**2005**). "Speaker normalization in speech perception," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell, Malden, MA) pp. 363–389.

Johnson, T. L., and Strange, W. (**1982**). "Perceptual constancy of vowels in rapid speech," J. Acoust. Soc. Am. **72**, 1761–1770.

Jongman, A., and Moore, C. (**2000**). "The role of language experience in speaker and rate normalization processes," Proceedings of the Sixth International Conference on Spoken Language Processing, I, pp. 62–65.

Joos, M. A. (**1948**). "Acoustic Phonetics," Language **24**, 1–136.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.

Leather, J. (**1983**). "Speaker normalization in perception of lexical tone," J. Phonetics **11**, 373–382.

Liberman, A. M., Cooper, F. S., Shankweiler, D. S., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psychol. Rev. **74**, 431–461.

Lin, T., and Wang, W. Y. (**1985**). "Tone perception," Chinese Linguistics J. **2**, 59–69 [in Chinese].

Marr, D. (**1982**). *Vision* (Freeman, New York).

Matthews, S., and Yip, V. (**1994**). *Cantonese: A Comprehensive Grammar* (Routledge, London).

Miller, J. L., and Dexter, E. R. (**1988**). "Effects of speaking rate and lexical status on phonetic perception," J. Exp. Psychol. Hum. Percept. Perform. **14**, 369–378.

Moore, C. B., and Jongman, A. (**1997**). "Speaker normalization in the perception of Mandarin Chinese tones," J. Acoust. Soc. Am. **102**, 1864–1877.

Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.

Newman, R. S., and Sawusch, J. R. (**1996**). "Perceptual normalization for speaking rate: Effects of temporal distance," Percept. Psychophys. **58**, 540–560.

Nusbaum, H., and Magnuson, J. (**1997**). "Talker normalization: Phonetic

constancy as a cognitive process," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix (Academic, San Diego), pp. 109–132.
Nusbaum, H. C., and Schwab, E. C. (**1986**). "The role of attention and active processing in speech perception," in *Speech Perception*, edited by E. C. Schwab and H. C. Nusbaum, Pattern Recognition by Human and Machines, Vol. **1** (Academic, San Diego), pp. 113–157.
Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**, 355–376.
Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker contingent process," Psychol. Sci. **5**, 42–46.
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
Pierrehumbert, J. A. (**1979**). "The perception of fundamental frequency declination," J. Acoust. Soc. Am. **66**, 363–369.
Rose, P. (**2000**). "Hong Kong Cantonese citation tone acoustics—A linguistic-tonetic study," in Proceedings of the Eighth Australian International Conference on Speech Science and Technology, edited by S. Barlow, Australian Speech Science and Technology Association, pp. 198–203.
Rose, P., July 18, 2005 (personal communication).
Singh, P. G., and Hirsh, I. J. (**1992**). "Influence of spectral locus and F0 changes on the pitch and timbre of complex tones," J. Acoust. Soc. Am. **92**, 2650–2661.
Stevens, K. N., and Halle, M. (**1964**). "Remarks on analysis by synthesis and distinctive features," in *Proceedings of the AFCRL Symposium on Models for the Perception of Speech and Visual Form*, edited by W. Wathen-Dunn (MIT, Cambridge), pp. 88–102.
Vance, T. J. (**1976**). "An experimental investigation of tone and intonation in Cantonese," Phonetica **33**, 368–392.
Verbrugge, R. R., Strange, W., Shankweiler, D. P., and Edman, T. R. (**1976**). "What information enables a listener to map a talker's vowel space?," J. Acoust. Soc. Am. **60**, 198–212.
Whalen, D. H., and Xu, Y. (**1992**). "Information for Mandarin tones in the amplitude contour and in brief segments," Phonetica **49**, 25–47.
Wong, P. C. M. (1998). "Speaker normalization in the perception of Cantonese level tones," Master's thesis, University of Texas at Austin (unpublished).
Wong, P. C. M., and Diehl, R. L. (**2003**). "Perceptual normalization for inter- and intratalker variation in Cantonese level tones," J. Speech Lang. Hear. Res. **46**, 413–421.

The relative roles of vowels and consonants in discriminating talker identity versus word meaning[a)]

Michael J. Owren[b)] and Gina C. Cardillo[c)]
Department of Psychology, Cornell University, Ithaca, New York 14853

(Received 6 May 2005; revised 29 November 2005; accepted 30 November 2005)

Three experiments tested the hypothesis that vowels play a disproportionate role in hearing talker identity, while consonants are more important in perceiving word meaning. In each study, listeners heard 128 stimuli consisting of two different words. Stimuli were balanced for same/different meaning, same/different talker, and male/female talker. The first word in each was intact, while the second was either intact (Experiment 1), or had vowels ("Consonants-Only") or consonants ("Vowels-Only") replaced by silence (Experiments 2, 3). Different listeners performed a same/different judgment of either talker identity (Talker) or word meaning (Meaning). Baseline testing in Experiment 1 showed above-chance performance in both, with greater accuracy for Meaning. In Experiment 2, Talker identity was more accurately judged from Vowels-Only stimuli, with modestly better overall Meaning performance with Consonants-Only stimuli. However, performance with vowel-initial Vowels-Only stimuli in particular was most accurate of all. Editing Vowels-Only stimuli further in Experiment 3 had no effect on Talker discrimination, while dramatically reducing accuracy in the Meaning condition, including both vowel-initial and consonant-initial Vowels-Only stimuli. Overall, results confirmed *a priori* predictions, but are largely inconsistent with recent tests of vowels and consonants in sentence comprehension. These discrepancies and possible implications for the evolutionary origins of speech are discussed. [DOI: 10.1121/1.2161431]

PACS number(s): 43.71.Es, 43.71.Bp [RLD] Pages: 1727–1739

I. INTRODUCTION

Although there is great variation among the sounds found in the thousands of languages used around the world, one of the most fundamental distinctions that can be drawn is between vowels and consonants (e.g., Ladefoged, 2001; Ladefoged and Maddieson, 1996). Both are basic components of speech, yet regularly differ in continuity, voicing, formant structure, and perceptual quality. While each kind of sound is used to form linguistic contrasts, vowel sounds are also particularly rich in indexical cues—acoustic features that are correlated with personal characteristics such as the biological sex, individual identity, age, and emotional state of the vocalizer. However, humans are not the only animals that convey indexical cues through vocalization. Taking a comparative perspective, it is important that whereas vowel-like sounds are routine in the vocal repertoires of monkeys, apes, and other mammals, consonant-like features are virtually absent (e.g., Lieberman, 1975). One implication is that vowel-like-calls were likely a central component of the nonhuman-primate-like vocal repertoire of the earliest bipedal apes (known as *hominins*, Berger, 2001), and that speech evolution was importantly associated with the eventual emergence of articulatory modulation of these sounds to produce consonantal contrasts. If so, the differentiated evolutionary origins of vowels and consonants should also be apparent in vestigial form in modern speech. The current work therefore tested the hypothesis that listeners hearing individual words are better able to discriminate talker identity from vowels, whereas meaning discrimination is better with consonants.

A. Evidence of fundamental vowel-consonant differences

Studies of phonological perception, word comprehension, and speech acoustics have consistently indicated that consonants and vowels play divergent roles in speech communication (see Mirman *et al.*, 2004 for a recent review). For example, whereas categorical perception can be readily demonstrated with stop consonants (e.g., Liberman *et al.*, 1957), outcomes are at best equivocal for vowels (e.g., Liberman *et al.*, 1967; Repp, 1984). Fujisaki and Kawashima (1970) suggested that this difference reflects better auditory memory for vowels, which are therefore less subject to label-based categorization than are consonants (see also Eimas, 1963; Pisoni, 1973, 1975; Healey and Repp, 1982). An implication of such findings is that there are separate memory stores for vowels and consonants (e.g., Cole, 1973; Pisoni, 1973), although this conclusion is not universally embraced (e.g., Ladefoged, 2001; Monaghan and Shillcock, 2003). Ades (1977) has for instance proposed that a vowel routinely spans a larger psychophysical range than does a consonant, which has functional implications for discrimination, identification, and memory performance (but see Macmillan *et al.*, 1988).

[a)]Portions of this work were presented at the 143rd meeting of the Acoustical Society of America, Pittsburgh, PA, June 2002.
[b)]Present address: Department of Psychology, Georgia State University, P.O. Box 5010, Atlanta, GA 30302-5010; electronic mail: owren@gsu.edu
[c)]Present address: Department of Speech and Hearing Sciences, University of Washington, Box 354875, Seattle, WA 98105-6246; electronic mail: ginacc@u.washington.edu

There is also evidence of differences in underlying neural control of vowels and consonants, both in production and perception. In one example, Caramazza *et al.* (2000) have provided case-study data indicative of a double dissociation between neural mechanisms involved in producing the two sound-types (also see Miceli *et al.*, 2004). Other evidence shows that people with aphasia exhibit more errors in producing consonants than vowels (reviewed by Monaghan and Shillcock, 2003). Even in neurologically intact talkers, consonants are typically associated with more production errors than are vowels (Fromkin, 1971), including being inadvertently omitted (Su, 1999). Consonants are likely to be misperceived as well, which can affect meaning interpretation (Bond and Garnes, 1980). Differences also appear in hemispheric laterality effects, which are prominent and robust when listeners discriminate consonant-vowel (CV) syllables, but become weak and inconsistent when only vowels are presented (e.g., Shankweiler and Studdert-Kennedy, 1967; Cutting, 1974; Jäncke *et al.*, 2002; see also Diesch and Luce, 1997). Direct electrical stimulation of language-related cortical regions (i.e., left superior temporal gyrus) is requisitely disruptive to consonant perception, while leaving vowel perception undisturbed (Boatman *et al.*, 1997).

Hearing loss can have differential effects on vowel and consonant perception, depending on the type of impairment and treatment involved. People with sensorineural loss that leaves them less able to perceive higher frequency energy typically show better perceptual performance with vowels than with consonants (Boothroyd, 1984; Dubno *et al.*, 1982; Lindholm *et al.*, 1988; Zeng and Turner, 1990). However, hearing-impaired listeners using cochlear implants tend to perceive consonants more accurately than vowels (Fu and Shannon, 1998; Teoh *et al.*, 2003; Välimaa *et al.*, 2002). This difference likely reflects the fact that cochlear implants preserve the temporal information required for consonant recognition, but do not provide sufficient spectral resolution for consistently accurate vowel perception.

B. Indexical versus linguistic cuing

In addition to its linguistic significance, speech is richly endowed with cues to a talker's personal characteristics, including biological sex, individual identity, age, and emotional state (Abercrombie, 1967; Ladefoged and Broadbent, 1957). While indexical cues are not inherently separable from linguistic aspects (summarized by Lachs *et al.*, 2000), listeners are nonetheless typically able to easily identify the phonetic content of an utterance independent of the individual producing it. Available evidence suggests that in contrast to consonants, vowels are particularly rich in indexical cues, specifically due to their combination of a stable fundamental frequency (F_0), rich harmonic structure, and supralaryngeal filtering effects (reviewed by Bachorowski and Owren, 1999). In talker-sex discrimination, for example, Whiteside (1998a, 1998b) has reported that participants labeling 50- and 100-ms speech sounds as being from male versus female talkers were 98.9% correct for vowels, but only 64.4% correct for unvoiced fricatives (see also Lass *et al.*, 1976). Owren and Bachorowski (2006) have further shown that only a very brief vowel segment is needed in order for listeners to hear the talker's sex, finding a "threshold" of less than two phonation cycles (roughly 14 ms of male speech, and 8 ms of female speech).

C. Vowels and consonants in comparative context

Fundamental differences between vowels and consonants can also be expected from a comparative perspective, because while vowel-like sounds are widespread among mammals, functional use of consonant(-like) sounds appears to be exclusive to human speech. Lieberman (1975) was among the first to point out that nonhumans such as monkeys and apes lack many of the articulatory maneuvers that routinely occur in human speech. Some articulatory capabilities are present, for example in opening and closing the lips, raising and lowering the larynx, and opening and closing the mandible (e.g., Hauser *et al.*, 1993; Shipley *et al.*, 1991; although see Riede and Zuberbühler, 2003; Riede *et al.*, 2005). Overall, however, primates and other nonhuman mammals are "source-focused," with qualitatively different sounds first and foremost being associated with nonlinear laryngeal biomechanics rather than sound-shaping at supralaryngeal levels (e.g., Wilden *et al.*, 1998; Fitch *et al.*, 2002). In contrast, human speech is strongly "filter-focused," relying on controlled, dynamic modulation of the supralaryngeal cavities through movements of the tongue, lips, and jaw to produce linguistic contrasts. For humans, qualitative shifts in vocal-fold oscillation associated with nonlinear phenomena are considered pathological, interfering with effective speech communication rather than enhancing it (Titze *et al.*, 1993; reviewed by Owren, 2003).

However, when mammals such as nonhuman primates do produce harmonically structured calls, their sounds can be very similar to human vowels. One striking instance is the vowel-like "grunt" call of baboons (*Papio cynocephalus ursinus*; Owren *et al.*, 1997). Grunts produced by adult-female baboons are virtually indistinguishable from prototypical schwa sounds produced by adult male humans, showing similar frequency values both for F_0 and formants (Owren *et al.*, 1997; Rendall *et al.*, 2004). Formants are predominant among the features differentiating vocalizers, and wild baboons have in turn been found to discriminate among individuals based on grunt acoustics (Rendall *et al.*, 1999). In other studies, free-ranging rhesus macaques (*Macaca mulatta*) readily discriminate among callers from hearing their vowel-like "coo" vocalizations (Rendall *et al.*, 1996), again with formants being the predominant distinguishing feature (Rendall *et al.*, 1998). Vowel-like sounds from humankind's nearest biological relatives, the common chimpanzees (*Pan troglodytes*) and bonobos (*P. paniscus*), have not been tested in this way. Nonetheless, spectrographic evidence reveals that these species also produce these kinds of calls (e.g., Bauer, 1986; Goodall, 1986; de Waal, 1988; Bermajo and Omedes, 1999).

D. Expectations for indexical versus linguistic cuing in human speech

Comparative observations thus suggest that the evolution of linguistic speech was associated with the emergence of consonantal contrasts and the supralaryngeal articulatory maneuvers that underlie these sounds. If so, vowel sounds might be considered a kind of carrier signal that is modulated to produce linguistic content (e.g., Dudley, 1939; Traunmüller, 1994; Owren, 2003; although see MacNeilage, 1998; MacNeilage and Davis, 2000). Given the indexical richness of unarticulated, vowel-like sounds in nonhuman primates, analogous components of human vowels are expected to be similar. Conversely, it follows that temporal modulations imposed on those sounds to produce consonants should be primarily associated with linguistic content. The latter view is supported by evidence from orthography, including the effectiveness of notation systems such as written Hebrew and shorthand that omit vowels entirely.

Though there are relatively few published empirical data that directly compare the roles of consonants and vowels, recent studies by Cole *et al.* (1996) and by Burkle and colleagues (Burkle, 2004; Burkle *et al.*, 2004) have found that listeners show significantly better comprehension of sentence-length stimuli when hearing vowels-only rather than consonants-only versions. Listeners in each set of experiments heard sentences selected from the TIMIT corpus (Garofolo *et al.*, 1990), with vowels and consonants replaced by either noise or sinusoidal sounds. Replacement was based on segment boundaries provided with the database, and the participant's task in each study was to write down or repeat as many words as possible from each sentence. In both cases, a robust difference was found, with significantly more words being identified from vowel-based sentences than from consonant versions.

The experiments reported here resemble the studies of Cole *et al.* and Burkle and colleagues in testing for differential importance of vowels and consonants in speech communication, but with several critical differences. First, the work contrasted the roles of vowels and consonants in both indexical and linguistic realms, in line with the above-outlined comparative/evolutionary perspective. Second, the task used individual words rather than sentence-length stimuli, given that any of a number of sentence-level properties could provide indexical or linguistic cues, including overall pitch contour, stress patterns, durations of segments and words, and syntactic structure. While these kinds of cues are likely important in normative speech communication, the goal was to test for effects that might be closer to those occurring when supralaryngeal articulation of vowel-like sounds first appeared in an early hominin. Third, deleted segments were set to silence in order to decrease possible phonemic restoration effects that could result from being replaced by noise or other sounds (reviewed by Warren, 1999). A variety of evidence indicates that phonemic restoration involves the listener's deep-seated linguistic knowledge, with sentence-level stimuli encouraging top-down processing and isolated words bringing bottom-up processing more into play (Samuel, 1981). Finally, a same-different, signal-detection design was used so as to directly compare talker- and meaning-discrimination accuracy (Macmillan and Creelman, 1991).

Three experiments were conducted, each of which involved listeners tested either in talker-discrimination or meaning-discrimination conditions. This between-groups design avoided carry-over effects that could result from listeners hearing the stimuli more than once. The first study established baseline performance levels by testing listeners with intact versions of the stimulus words. The second and third experiments then contrasted the relative contributions of vowels and consonants in each condition by presenting either "Vowels-Only" or "Consonants-Only" versions of the target stimuli. To minimize any effects of the particular words being presented, half the listeners in each experiment heard a particular set of words with either vowels or consonants removed, while the remaining listeners heard the same set of words manipulated in converse fashion.

II. GENERAL METHODS

A. Subjects

Participants were university undergraduates who described themselves as native English speakers with no speech- or hearing-related difficulties. Stimuli were based on speech recorded from 8 males and 8 females between 19 and 24 years of age that served as talkers.

B. Materials and procedure

1. Stimuli

Talkers were recorded individually in a sound-treated room using a Crown CM-311A condenser head-worn microphone (Crown, Elkhart, IN) connected to an SGI O2 workstation (Silicon Graphics Inc., Mountain View, CA). An experimenter was present throughout to position the microphone, adjust the input level, and ensure that talkers spoke each item clearly and naturally. Stimulus words were digitized directly to disk at 22.05 kHz using the Praat program (Boersma and Weenink, 2005). Further stimulus preparation used both Praat and ESPS/waves+5.3 (Entropic Research, Washington, DC) software.

A total of 256 common English-language words were used in the stimuli, which consisted of pairs of words whose meanings were either obviously similar or obviously dissimilar. The stimuli included 50, 38, and 40 pairs of nouns, verbs, and adjectives, respectively. Half the resulting 128 stimuli thus included (near-) synonyms, with word meanings being unrelated in the other half. All words had been screened in unedited form by a group of 11 native English speakers, allowing only unambiguous items to be used. The final set of 128 word pairs were quasirandomly divided among the 16 talkers, who each recorded a unique set of 16 words from an alphabetized list. Each individual's recordings were used to create 8 same- and 8 different- meaning pairs. All words were then edited into separate files, rescaled to the full 16 -bit amplitude range available, and assembled as stimuli with the first word always ending at exactly 1000 ms and the second word beginning at 1700 ms (illustrated in Fig. 1). The second word was presented again beginning at 3000 ms, with zero-padding added as needed to both ends of the file.

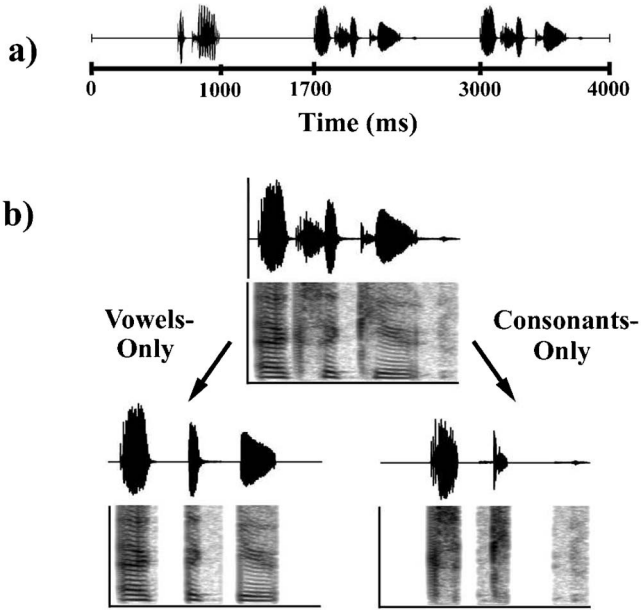


FIG. 1. (a) A schematic illustration of the stimuli presented on each trial. Stimuli consisted of an unedited target word and two instances of the comparison word, structured such that the target word terminated at the 1000-ms point and the first comparison word always began at 1700 ms. (b) Waveforms and narrowband spectrograms (22 050-Hz sampling rate, 0.05-ms Gaussian window) of the stimulus word "execute" spoken by a female talker, shown intact (top) and with either consonants (Vowels-Only) or vowels (Consonants-Only) replaced by silence (bottom). The visible frequency range is 5500 Hz.

The second word was presented twice in order to reduce the difficulty of the task for listeners hearing edited versions.

While all stimuli were intact in Experiment 1, Experiments 2 and 3 used versions in which the second word in each word-pair was edited so as to replace either all its consonants ("Vowels-Only" condition) or all its vowels ("Consonants-Only" condition) with silence (illustrated in Fig. 1). For editing purposes, vowels were defined as nonobstructed, voiced sounds, and included /i/, /ɪ/, /eɪ/, /ɛ/, /æ/, /a/, /ɔ/, /ʊ/, /u/, /ʌ/, and /aɪ/. Consonants were sounds whose articulation included a vocal-tract constriction at any point above the larynx, and included /p/, /t/, /k/, /b/, /d/, /g/, /m/, /n/, /ŋ;/, /f/, /v/, /θ/, /ð/, /s/, /z/, /ʃ/, /ʒ/, /h/, /r/, /l/, /j/, and /w/. Liquids (/r/, /l/) and glides (/y/, /w/) were avoided as possible during word selection, and occurred in only 18 of the 128 edited stimuli.

Word editing was done twice to create both a Vowels-Only and a Consonants-Only version of each (see Fig. 1). Editing was based on inspecting 25-ms sections of both wave form and wideband-spectrogram displays for evidence of vowels. Criteria for identifying a vowel included the presence of sustained voicing, clear formant structure, and no sign of vocal-tract constrictions. Only segments showing vowel characteristics were retained for Vowels-Only stimuli, while the same segments were eliminated from Consonants-Only stimuli. Adjacent phonemes often lack precise boundaries (e.g., Hardcastle and Hewlitt, 1999), which prompted a conservative editing strategy. Specifically, visible formant transitions occurring between vowel and consonant segments were excluded from both. The approach was thus to emphasize steady-state centers in Vowels-Only stimuli, while removing formants from all segments except voiced consonants occurring in the Consonants-Only stimuli. As a result, boundaries for adjacent segments could be slightly different in the two corresponding versions of a given stimulus word. To help avoid audible artifacts, cuts were always made at wave form zero-crossings. After editing, files were again amplitude-scaled and assembled as stimuli.

Forty-six of the altered words began with vowels, while 82 began with consonants. Fifty-eight were two syllables long, 44 were three syllables, 23 were four syllables, and 3 were five syllables. Each of the 128 stimuli was unique, and was heard only once so as to minimize learning effects. The stimulus set included equal numbers of male and female talkers, equal numbers of same and different talkers, and equal numbers of stimuli with same and different meanings.

2. Perceptual testing

Perceptual testing occurred in a small room with 5 booths equipped with Beyerdynamic 831 headphones (Beyerdynamic, Farmingdale, NY) and 4-button TDT response boxes (Tucker-Davis Technologies, Gainesville, FL). Booths were operated from an adjacent room using TDT modules, a computer, and custom-written software (B. Tice & T. Carrell, available from www.hush.unl.edu/LabResources.html). Stimuli were low-pass-filtered at 10 kHz and presented against a quiet background at a comfortable listening level (approximately 76–80 dB measured at the headphones). Listeners were tested in groups of up to four after completing an informed-consent form and a demographic questionnaire that also asked about speech- or hearing-related problems. They were quasirandomly assigned to condition, and were instructed that the task was to decide whether words heard on each trial had been produced by the same talker or two different talkers (Talker condition), or had the same or different meaning (Meaning condition). Responding required pressing one of two buttons, labeled either or "Same Speaker" versus "Different Speaker," or "Same Meaning" versus "Different Meaning." Sixteen practice trials were completed before testing began, with talkers and words that were different from those used as experimental stimuli. Listeners had 4600 ms in which to respond, there was a 2500-ms intertrial interval, and the entire session took approximately 30 min. Word pairs were presented in random order within each session, with button-label position being reversed on alternate testing days.

3. Data analyses

Listener responses were scored as hits (correctly selecting "Same"), misses (incorrectly selecting "Different"), correct rejections (correctly selecting "Different"), or false alarms (incorrectly selecting "Same"). Proportions of hits and proportions of false alarms were used to calculate d' values for each listener using the "differencing" model for same-different, signal-detection designs (Macmillan and Creelman, 1991). Statistical analyses were repeated-measures ANOVA comparisons conducted with NCSS software (Jerry Hintze, Kaysville, UT).

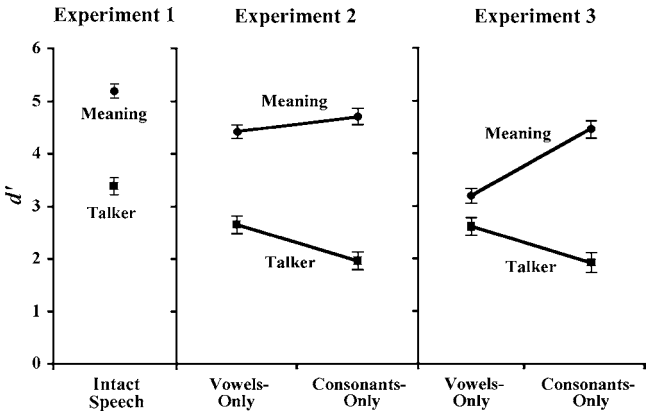


FIG. 2. Means and standard errors of d' results by condition in Experiments 1–3. Experiment 1 provided baseline performance with the stimuli in unedited form, Experiment 2 tested the effects of replacing either consonants (Vowels-Only) or vowels (Consonants-Only) with silence, and Experiment 3 replicated Experiment 2 after further editing of Vowels-Only stimuli by 50% to reduce possible coarticulation effects.

III. EXPERIMENT 1: BASELINE PERFORMANCE FOR IDENTITY AND MEANING

The goal of the first experiment was to establish baseline performance levels for judging talker identity and word meaning from a common stimulus set. The stimuli were designed such that listeners could be asked to make same-different judgments about either aspect, producing d' values that could then be compared across the two conditions.

A. Subjects

Twenty-four participants (23 females) were tested as listeners, with 12 individuals quasirandomly assigned to each of the Talker and Meaning conditions. All received extra credit in an undergraduate class. Data were excluded from an additional 7 listeners, 5 of whom were not native-English speakers, and 2 due to equipment malfunction.

B. Materials and procedure

Listeners were tested with intact versions of the stimuli.

C. Results and discussion

Results for mean d' in both Talker and Meaning conditions are illustrated in Fig. 2 (left panel). Repeated-measures ANOVA revealed a significant main effect of Condition, with overall values in the Meaning condition ($M=5.2$, SE=0.1) being higher than in the Talker condition ($M=3.2$, SE=0.2), $F(1,22)=63.6$, $p<0.001$. The latter result was not predicted, as discriminating talker identity was arguably the simpler task. Although listeners were likely attending to a variety of acoustic cues in making this judgment, the task only required listeners to decide whether two voices were sufficiently similar to have come from the same individual. In contrast, attending to meaning involved at least three separable components, namely linguistic decoding, retrieval of word meanings, and deciding whether the two meanings were the same. This result is considered further in Sec. VI.

Analyzing results in the Talker condition separately revealed no effect of Talker sex, but did show a weak interaction between Talker sex and same-different Talker, $F(1,11)=5.86$, $p<0.05$. Specifically, listeners performed better with male sounds when the two words in a stimulus pair had different meanings, but did best with female talkers when word-pair meanings were the same. There was a main effect of Talker sex in the Meaning condition, with listeners showing higher d' values for stimulus words produced by female talkers ($M=6.2$, SE=0.17) compared to male talkers ($M=4.2$, SE=0.14), $F(1,11)=83.5$, $p<0.0001$. There was again an interaction between Talker sex and same-different Talker, with listeners showing greater accuracy with female sounds when word-pairs were from two different individuals, but performing better with male sounds when both words were from the same talker, $F(1,11)=7.86$, $p<0.05$. These outcomes are also further discussed in Sec. VI.

IV. EXPERIMENT 2: VOWELS AND CONSONANTS IN IDENTITY AND MEANING

The second experiment was designed to address the central hypothesis of the work, namely that vowels play a more important role than consonants in hearing talker identity, while consonants are more important than vowels in understanding word meaning. The stimuli used in Experiment 1 were therefore edited so as to set either vowels or consonants to silence, as described in Sec. II. Following Cole *et al.* (1996), it was expected that segments in the Vowels-Only stimuli could retain some cues to adjoining consonant sounds even after editing. Vowels and consonants were nonetheless predicted to show differential contributions to talker and meaning discrimination.

A. Subjects

Forty participants (27 females) served as listeners, with 20 individuals quasirandomly assigned to each of the Talker and Meaning conditions. Results from 5 additional listeners were excluded. Two were eliminated because they were not native-English speakers, 2 due to equipment malfunction, and 1 because she recognized the voice of one of the talkers.

B. Materials and procedure

Listeners were tested as in Experiment 1, except that they heard edited versions of the stimuli, were informed that the stimuli would include edited words, and warmed up with comparably edited versions during practice trials. In addition to the previous counterbalancing measures, listeners were tested with one of two complementary and mutually exclusive stimulus lists (Lists "A" and "B"). In each list, half of the stimuli were Vowels-Only and half were Consonants-Only, and the two lists reflected converse manipulations for each edited word. Ten listeners in each condition heard each list, with a total of 3 males and 17 females hearing List A, and 8 males and 12 females hearing List B. This counterbalancing was designed to reduce the possibility that the particular words used could differentially affect listener performance.

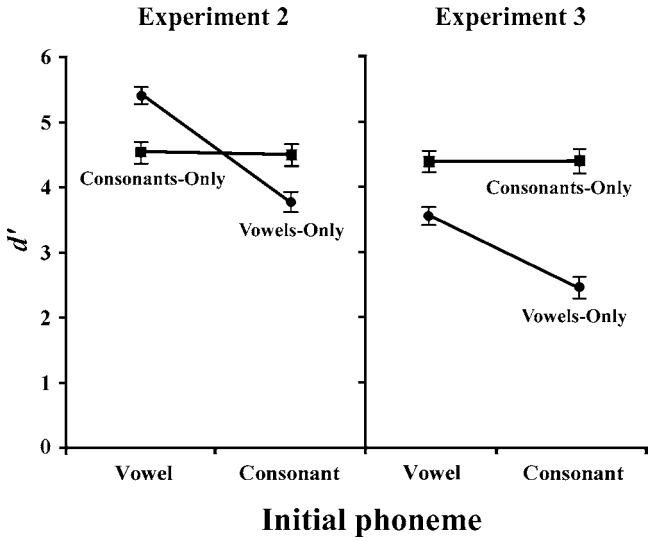


FIG. 3. Means and standard errors of d' results in Experiments 2 and 3, shown according to Meaning vs Talker condition and whether the first letter in the original stimulus word was a vowel or consonant (Initial-Phoneme).

C. Results and discussion

Listener responses were scored and analyzed as before, with mean d' outcomes for both Talker and Meaning conditions shown in Fig. 2 (middle panel). There were no overall effects of either List or Listener sex, and these factors are not considered further. As in Experiment 1, there was a significant main effect of Condition, with d' values being higher in the Meaning condition ($M=4.7$, SE=0.10) than in the Talker condition ($M=2.4$, SE=0.12), $F(1,38)=88.3$, $p<0.0001$. However, both values had also declined relative to corresponding outcomes with unedited stimuli.

1. Vowels versus consonants

The result of greatest interest was that while listeners showed no main effect of Vowels-Only versus Consonants-Only, $F(1,38)=2.61$ ns, this factor interacted significantly with Talker versus Meaning condition, $F(1,38)=14.8$, $p<0.001$. As illustrated in Fig. 2, listeners exhibited higher d' values for Vowels-Only stimuli in the Talker condition, while performing better with Consonants-Only stimuli in the Meaning condition. These outcomes also appeared as main effects when tested separately within condition, with discrimination in the Talker condition being significantly better with Vowels-Only than with Consonants-Only stimuli, $F(1,19)=9.59$, $p<0.01$, and discrimination in the Meaning condition being significantly better with Consonants-Only stimuli than Vowels-Only stimuli, $F(1,19)=5.57$, $p<0.05$.

A follow-up analysis was also conducted in the Meaning condition to examine whether the initial phoneme in each word affected performance, given the importance that initial sounds can play in lexical access (e.g., Marslen-Wilson, 1987). As the stimulus set included 36 more consonant-initial than vowel-initial words, this comparison was conducted to check whether discrimination performance was particularly affected by the presence or absence of the initial sound. Results are shown in Fig. 3 (left panel), and revealed both a main effect of Initial-Phoneme, $F(1,19)=17.2$, $p<0.001$, and an interaction with editing condition, $F(1,19)=10.9$, $p<0.01$. On the one hand, performance with Consonants-Only stimuli was unaffected by whether the initial phoneme was present (as a consonant) or absent (as a missing vowel). On the other hand, an unexpected outcome appeared, with listeners showing the best performance of any condition when hearing Vowels-Only stimuli that began with a vowel. Performance was clearly worst with Vowels-Only stimuli in which an initial consonant had been removed from the edited word. Taken together then, the two produced a mean outcome that fell significantly below performance with Consonants-Only stimuli. These results are further discussed in the following.

2. Other comparisons

There was a main effect of Talker sex in the Talker condition, with listeners showing better overall performance hearing males than when hearing females, $F(1,19)=9.59$, $p<0.01$. While there was no main effect of same-different Meaning in this condition, there was an interaction between Talker sex and Meaning, $F(1,19)=4.41$, $p<0.05$. This effect reflected that it was easier to distinguish among male talkers producing words with different meanings. Another significant interaction was that female talkers were found to be particularly difficult to discriminate from Consonants-Only stimuli, $F(1,19)=4.39$, $p<0.05$. In the Meaning condition, there was no main effect of Talker sex, but there was a main effect of same-different Talker, $F(1,19)=12.27$, $p<0.01$. Here, performance was higher when two different talkers produced the words than when both were from the same talker. Performance was lowest when both words had been produced by the same male talker, $F(1,19)=12.74$, $p<0.01$.

3. Summary and conclusions

Results of this experiment were consistent with expectations in some respects, while also producing an unexpected outcome. Listeners were clearly better able to hear talker identity from vowels than from consonants, and were better overall in discerning word meaning from consonants than from vowels. Follow-up analyses in the Meaning condition were consistent with this overall result in showing that it made no difference for Consonants-Only stimuli whether the edited word had originally begun with a vowel or consonant. However, the best performance of all was actually found to occur with Vowels-Only words that also began with a vowel, with the worst performance occurring with Vowels-Only words that were missing an initial consonant. Results with vowel-initial Vowels-Only stimuli were thus contrary to our hypotheses and consistent with findings from Cole *et al.* and Burkle and colleagues. However, they were also different from results with consonant-initial Vowels-Only versions, suggesting that an initial vowel can be a disproportionately informative event, particularly in comparison to later vowels in a word. Finding no effect whatsoever of initial phoneme with Consonants-Only stimuli underscores the possibility that consonants play a critical role in linguistic cuing no matter where in a word they appear.

One possibility to consider was therefore whether initial vowels could be differentially important simply because there are fewer words that begin with vowels than with consonants. If so, then the degree of uncertainty reduction resulting from hearing even compromised vowel-related information in the initial position could have a disproportionate effect on Meaning performance. This possibility was investigated by examining the relative frequency of initial phonemes in the MRC Psycholinguistic Database (Wilson, 1988; available at http://www.psy.uwa.edu.au/mrcdatabase/mrc2.html), which is an Australian-English database that includes 150 837 entries. The search was set for all content words, meaning nouns, verbs, and adjectives as were used here. Proper nouns were excluded. The result was a total of 7600 vowel- and diphthong-initial words, but 24 828 consonant-initial items. Searching somewhat more conservatively by excluding /l/, /r/, /y/, and /w/ as initial phonemes 21 553 consonant-initial entries. Based on these estimates then, there are only one-third to one-half as many vowel-initial as consonant-initial words in English, suggesting a significantly higher value for linguistically related information concerning a vowel sound that begins a word.

While the differential effects of editing vowels versus consonants were unmistakable in the Talker condition, the difference was thus both quite modest and more complicated in the Meaning condition. One factor likely to be important to both aspects is that the Vowels-Only stimuli retained coarticulation effects created by adjacent consonants in the original words. If so, these cues could have continued to provide linguistically relevant information to listeners in spite of the editing that had been done. For example, while editing was designed to also remove format transitions, any residual transition information could be a prominent source of information concerning preceding and following consonants (e.g., Liberman *et al.*, 1954). Coarticulation effects could also be affecting vowels beyond their immediate onset and offset components, with both preceding and anticipated consonantal constrictions influencing details of vowel articulation. The salience of anticipatory effects would arguably be particularly prominent in initial vowels found among the experimental stimuli, due to the stress placed on these sounds by talkers who were producing the words in isolation. The possibility of coarticulation effects was addressed in Experiment 3, which repeated the testing after further editing of the Vowels-Only stimuli that was designed to remove any residual format transitions and reduce the importance of other possible coarticulation effects. This replication also provided an opportunity to test the reliability of other statistically significant main effects and interactions.

V. EXPERIMENT 3: DISCRIMINATION WITH SHORTENED VOWEL SEGMENTS

The third experiment was motivated by the possibility that coarticulation-related cuing present in vowels in the Vowels-Only sounds had significantly affected Meaning discrimination outcomes in Experiment 2. One possible approach to eliminating such cues was to examine each vowel individually and to remove any visible coarticulation effects. However, that proved difficult and likely unreliable, prompting an alternative strategy of simply setting an additional 50% of each vowel in the Vowels-Only stimuli to silence. The rationale was that if consonant-related coarticulation effects created an important source of linguistic cuing in the Vowels-Only stimuli, removing material at the onset and offset of each vowel should have a disproportionate and deleterious impact on listener accuracy in the Meaning condition compared to the Talker condition. On the other hand, if steady-state components of vowels were cuing both talker and linguistic aspects, any one of a number of outcomes might occur. Mentioning just a few, there might be no change from results of Experiment 2, which would suggest that vowels do play a critical role in linguistic cuing even in isolated words. This outcome would be at least partially consistent with findings reported by Cole *et al.* and Burkle and colleagues, with the caveat that vowel-position within a word is an important factor. Second, the results could become more consistent with these previous findings, for instance if the advantage for initial-vowel Vowels-only stimuli proved replicable, but with the observed disadvantage for other Vowels-Only disappeared. Third, performance with Vowels-Only sounds could show a proportional decline in both the Talker and Meaning conditions due to floor effects created by removing so much of the sound segments remaining in these stimuli. Finally, further editing of the Vowels-Only stimuli could be found to have little effect on Talker performance, while noticeably and uniformly reducing discrimination accuracy in the Meaning condition. This last outcome would be the most consistent with the hypothesized differentiation of vowels and consonants in indexical and linguistic cuing.

A. Subjects

Forty participants (37 females) served as listeners, with 20 individuals quasirandomly assigned to each of the Talker and Meaning conditions. Results from 4 additional listeners were excluded. Two were eliminated because they were not native English-speakers, and 2 due to equipment malfunction and experimenter error.

B. Materials and procedure

The stimuli were the same as in Experiment 2, except that all vowels comprising comparison words in the Vowels-Only condition were reduced in duration by 50%. In each case, an equal amount was removed from the beginning and end of each segment, and was replaced by a corresponding amount of silence. Listeners were tested as in Experiment 2, except that they heard the reduced versions of Vowels-Only stimuli in both practice and experimental components. In this experiment, 20 listeners (3 males and 17 females) heard List A, while 20 listeners (all females) heard List B.

C. Results and discussion

Listener responses were scored, tabulated, and analyzed as in Experiment 2, with mean d' outcomes for both Talker and Meaning conditions shown in Fig. 2 (right panel). There was no overall effect of List and this factor is not considered further. Listener sex could not be tested, due to the small number of male listeners involved. Here again, there was a

main effect of Condition, with d' values being significantly higher in the Meaning condition ($M=3.9$, SE=0.12) than in the Talker condition ($M=2.1$, SE=0.13), $F(1,38)=50.1$, $p<0.0001$. However, both values were somewhat lower than in Experiment 2, presumably due to the additional editing of the vowel stimuli.

1. Vowels versus consonants

Unlike Experiment 2, listeners now showed a main effect of Vowels-Only versus Consonants-Only stimuli, showing modestly but significantly greater accuracy with the latter $F(1,38)=4.12$, $p<0.05$. However, this change was traceable to a substantial decline in Meaning discrimination with Vowels-Only stimuli, with Vowels-Only versus Consonants-Only interacting strongly with Talker versus Meaning condition, $F(1,38)=68.6$, $p<0.0001$. As illustrated in the figure, listeners again showed higher d' values for Vowels-Only stimuli in the Talker condition, while performing better with Consonants-Only stimuli in the Meaning condition. The additional editing of Vowels-Only stimuli had no evident effect on Talker discrimination, whereas Meaning discrimination values declined sharply. Consonants-Only stimuli were the same as those used earlier, and mean d' outcomes for these sounds were almost identical to those in the previous experiment. Separate analyses within each condition revealed strong corresponding main effects. In the Talker condition, Vowels-Only stimuli produced significantly better performance than Consonants-Only stimuli, $F(1,19)=21.2$, $p<0.001$. The advantage for Consonants-Only stimuli in the Meaning condition was also significant, and had become much stronger than in Experiment 2, $F(1,19)=49.4$, $p<0.0001$.

Results in the Meaning condition are shown by Initial-Phoneme in Fig. 3 (right panel). As in Experiment 2, there was both a main effect of this factor, $F(1,19)=5.7$, $p<0.05$, and an interaction with editing condition, $F(1,19)=11.8$, $p<0.01$. There was again no discernible effect of Initial-Phoneme with Consonants-Only stimuli, where performance was virtually identical to earlier results. For Vowels-Only stimuli, however, accuracy declined uniformly in the two Initial-Vowel conditions, thereby accounting for the dramatic overall decline in Meaning performance found with these sounds relative to the earlier experiment.

2. Other comparisons

Testing within the Talker condition revealed no main effects of either Talker sex, $F(1,19)=0.34$, or same-different Talker, $F(1,19)=2.23$, and there were no significant interactions. Performance in the Meaning condition also revealed no main effect of Talker sex, $F(1,19)=0.08$, *ns*, and the main effect of the Talker variable observed in Experiment 2 was now absent, $F(1,19)=1.90$, *ns*. However, there were two factors that produced significant interactions. For stimuli by male talkers, different-talker stimuli were easier to discriminate than were same-talker stimuli $F(1,19)=7.12$, $p<0.05$. In addition, whereas Meaning discrimination was slightly lower for male than for female Vowels-Only stimuli, the converse was observed for Consonants-Only stimuli, $F(1,19)=5.65$, $p<0.05$.

3. Summary and conclusions

Overall, the results confirmed and strengthened the conclusion that vowels were playing a more prominent role than consonants in talker discrimination, while consonants were more important than vowels in meaning discrimination. Removing initial and final portions of each Vowels-Only stimulus clearly lowered the accuracy of Meaning discrimination, while having no discernible effect on performance in the Talker condition. This outcome appears to confirm that the Vowels-Only stimuli used in Experiment 2 included critical coarticulatory effects of adjacent consonants and that these cues contributed disproportionately to Meaning judgments. The results also confirmed both the absence of an initial-position effect in Consonants-Only stimuli, and that initial vowels were playing a critical role in the Vowels-Only stimuli.

Most other effects proved to be inconsistent across Experiments 2 and 3. For example, main effects of Talker sex on Talker discrimination and of same-different Talker on Meaning discrimination both disappeared in Experiment 3. Two of three interaction effects from Experiment 2 were also absent, with Experiment 3 generating one interaction that had not been observed earlier. Only a single ancillary effect appeared in both studies, namely that Meaning discrimination was somewhat better when stimuli included words produced by two different males than when both words were from the same male. While consistent, this outcome was also modest, and likely not particularly important to the overall implications of the two experiments.

VI. GENERAL DISCUSSION

Taken together, results of the three experiments indicate that vowels and consonants play different roles when listeners attend to talker identity versus word meaning. Independent of talker or listener sex, Vowels-Only stimuli allowed more accurate Talker discrimination, whereas Consonants-Only stimuli allowed more accurate Meaning discrimination. A variety of additional effects occurred, two of which were stable and important. One was an overall performance difference in hearing talker versus hearing meaning, and the other was better meaning-discrimination with Vowels-Only stimuli that also began with a vowel. Various other differences that were primarily statistically significant interactions were more subtle and unstable, and appear relatively unimportant to the current work.

A. Additional outcomes

1. Hearing talker versus hearing meaning

Though not directly related to the main research question, a consistent finding across all three experiments was that overall performance in the Talker condition was significantly lower than in the Meaning condition. This outcome occurred in spite of the latter arguably being the more cognitively demanding task, involving high-level processes such

as linguistic decoding, retrieval of word meanings, and comparing these meanings for two different words. One interpretation of this difference is that listeners in both conditions were likely experiencing some interference in task-performance due to the irrelevant variation occurring on the unattended stimulus dimension (e.g., Garner, 1970). If so, this Garner interference effect should arguably be significantly stronger in the Talker condition due to the highly automatic nature of semantic processing (e.g., Neely, 1991). Variation in word meaning would therefore be expected to have a more deleterious effect on Talker performance than variation in talker identity would have on Meaning accuracy.

This interpretation leads to the prediction that using nonsense words instead of lexical items in the Talker condition should significantly increase performance accuracy relative to the Meaning condition. However, the results of Experiment 3 argue against this possibility, in that reducing linguistic cues by further editing of the Vowels-Only stimuli did not produce a requisite increase in Talker discrimination relative to that in Experiment 2. However, it is also possible that the additional editing had offsetting effects in the Talker condition. Shortening the vowel segments could have depressed performance with Vowels-Only stimuli by reducing indexical as well as linguistic cuing, while simultaneously boosting performance with these sounds by decreasing interference due to the irrelevant meaning dimension. Another explanation is that Talker discrimination was the more difficult task because the listeners were unfamiliar with the particular voices heard in the experiment. If so, significantly improved performance would be expected even for lexically meaningful items when spoken by individuals whose voices were familiar to the listeners.

2. Initial phoneme effects

Results in Experiments 2 and 3 also revealed a differential effect of whether the first phoneme in a stimulus was a vowel or a consonant, and whether this initial sound appeared in the edited stimulus. As illustrated in Fig. 3, both experiments showed that whereas retaining the initial phoneme has a clear influence on hearing meaning solely from vowel components, there is no analogous effect for consonant-based stimuli. These findings underscore that for word-level stimuli at least, linguistic cuing is more ephemeral in vowels than in consonants. Furthermore, the Initial-Phoneme effect observed in the Vowels-Only condition was robust to the additional editing of vowel segments occurring in Experiment 3. Although performance in the Vowels-Only condition dropped precipitously in that study, words with an initial vowel were still understood better than the other stimuli. This outcome may be best interpretable based on the large evident disparity between the number of English-language content words that begin with consonants versus vowels. In other words, that the ability to hear meaning from vowels alone is largely a function of the extent to which information about the initial phoneme can be used to limit the number of lexical possibilities.

The absence of an effect for Consonants-Only stimuli is also worth noting, as a variety of evidence now exists that selecting among lexicon items when hearing or reading individual words is an on-line process in which early phonemes are used to limit the search (e.g., Marslen-Wilson, 1987; Eberhard *et al.*, 1995; Miller and Eimas, 1995). An Initial-Phoneme effect would therefore arguably be expected for Consonants-Only stimuli as well. Its absence may simply reflect the difficulty of the task used here, in which so little contextual information was provided. Studies of listener processing of ambiguous words or sentences are typically less challenging, in that they focus on contiguous speech segments, rather than presenting stimuli consisting of noncontiguous fragments such as those used here. If this interpretation is correct, simplifying the task by limiting the number of possibilities for consonant-initial words should reveal an effect of the first phoneme for those stimuli as well.

3. Other effects

An effect noted in baseline testing in Experiment 1 was that listeners performed better in the Meaning condition when hearing female speech than when hearing male speech. This outcome is consistent with long-established findings that females are on average both faster and more accurate in speech articulation (e.g., Kimura, 1983; Hyde and Linn, 1988). In other words, while performance with the intact stimuli was high for both male and female talkers, words from the latter were likely somewhat more precisely spoken, somewhat easier to understand, and therefore also more accurately judged in the Meaning condition. However, Talker sex effects noted in Experiments 2 and 3 were found to be small and inconsistent. Any differences between the male and female tokens used in these studies therefore appear not to be important to the main findings.

B. Comparisons to previous studies

Finding that listeners were better able to hear word meanings from consonants than from vowels is consistent with the original expectations of Cole *et al.* (1996), but is largely inconsistent with both their findings and those of Brukle and colleagues (2004; Burkle *et al.*, 2004). In each of these studies, listeners showed clearly better word comprehension with vowels-only than with consonants-only stimuli. Several factors may have contributed to this discrepancy, with three in particular standing out.

First, both Cole *et al.* and Burkle and colleagues presented edited sentences in their experiments, thereby necessarily including information at the prosodic and syntactic levels that the word-based stimuli used here did not provide. Features such as overarching pitch contour, relative segment durations, stress patterns, and syntactic structure could all give listeners important clues to lexical identity. Cole *et al.* (1996) included an experiment designed to at least partially counteract such effects by reducing or enlarging all stimulus segments by 10 ms at both ends. While these manipulations did influence performance somewhat, listeners nonetheless continued to show significantly better comprehension of vowels-only than consonants-only sentences. However, running speech is known to include coarticulatory effects that occur not only among phonemes within words, but also across phrases. West (1999) suggests that these kinds of

phrase-level effects function to increase communicative efficiency by making the individual units of an utterance more cohesive and meaningful. Although it is not known whether coarticulation has a differential impact on vowels versus consonants, the addition of phrase-level effects to other suprasegmental cues may be crucial to the different outcomes observed at sentence-versus word-level testing.

Second, whereas vowels and consonants were differentially set to silence in the current experiment, both Cole *et al.* and Burkle and colleagues replaced excised segments with white noise, pink noise, or sinusoids. Phonemic restoration experiments suggest that this difference may have been important to comprehension outcomes (reviewed by Warren, 1999), particularly when contextually rich, sentence-level stimuli are involved. When a phoneme is replaced by noise, for example, it is common for listeners to report being unaware that anything is amiss.

Instead, listeners often describe an illusion of hearing the missing segment, identify that segment accurately, perceive the noise to be extraneous, and fail when trying to pinpoint where in the flow of words the noise occurred. In contrast, when comparable experiments are conducted with target phonemes replaced by silence, listeners readily hear the gaps, do not experience the illusory restoration effect, and can accurately report where the silences occurred (Bashford and Warren, 1987; Bashford *et al.*, 1992). This contrast underscores the top-down nature of the phoneme restoration effect (Repp, 1992), with listeners utilizing contextual information in the speech stream in light of a rich body of implicit linguistic knowledge. The size of the missing segments that can be tolerated thus becomes requisitely dependent on how much contextual information is available.

Another potentially significant factor is that by inserting noise, the previous studies may have created forward and backward masking effects that were most prominent for consonant-based stimuli. Parker and Diehl (1984), for example, asked listeners to identify vowels in a CVC context that had been replaced to varying degrees either by noise or by silence (further discussed in the following). Performance was virtually identical in the two conditions after replacing as much as 80% of the middle portion of the original vowel, but at 90% replacement error rates became significantly higher with noise-based than with silence-based insertions. This study was of course different in presenting syllable-, rather than word- or sentence-length stimuli, and in asking listeners to identify the vowel itself rather than surrounding consonants. However, as both Cole *et al.* and Burkle and colleagues were using noise to replace vowels in their entirety in consonants-only stimuli in their entirety, masking may have affected those surrounding components as well.

On the other hand, the use of silence in the current experiments may itself have had unintended deleterious consequences by misleading those listeners who were attempting to hear the sounds as meaningful speech rather than simply attending to cues to individual identity. Silent gaps can in and of themselves be communicatively significant, occurring for example in voiceless stop consonants, at the end of breath groups, to create emphasis, and as inadvertent hesitations. As a result, the occurrence of silence instead of noise could have been a source of misinformation rather than simply removing linguistic information. If so, listener perception of inserted silence as a communicatively significant event would have depressed performance in the Meaning condition, and would likely be more disruptive when listening to vowels-only than to consonants-only words.

Third, the sound editing done in the current work was more draconian in its treatment of consonant-vowel boundaries than in the previous studies. Both Cole *et al.* and Burkle and colleagues report that editing was based on phonemic boundaries provided with the TIMIT database, which would have results in better preservation of dynamic, transitional elements of the original speech. In the current work, formant transitions in particular were eliminated as much as possible from either vowel or consonant components. To the extent that components such as formant transitions differentially appeared in vowel segments, the perceptual task presented to listeners by Cole *et al.* and Burkle and colleagues would have been significantly different from the one used here.

C. The critical role of temporal modulation in linguistic speech

1. Silent-center vowels

A theme that has been touched on only tangentially to this point is the critical importance of temporal modulation in producing linguistic content in speech. Consistent with the general perspective that speech evolved through imposing articulatory gestures on a vowel-based carrier signal (Dudley, 1939; Traunmüller, 1994, Owren, 2003), a variety of data from linguistic processing point to the critical role of the dynamic properties of speech. One kind of evidence comes from "silent-center" vowel experiments, in which listeners have been found to be surprisingly accurate in correctly identifying vowels embedded in a CVC context even when the center portion of the vowel segment is set to silence (e.g., Parker and Diehl, 1984; Strange, 1989a, 1989b; Jenkins *et al.*, 1994). In fact, vowel perception under these conditions is more accurate than when vowel-centers are presented in isolation (Strange, *et al.*, 1976; Verbrugge and Rakerd, 1986). Hillenbrand and Nearey (1999) have further shown that whereas synthetic vowels generated with steady-state formants are correctly identified about 75% of the time, stimuli that track naturally occurring formants produce approximately 90% correct identification. According to these studies, it is the dynamic, time-varying onsets and offsets of vowels that are critical to correct phonetic identification. Although cues from more static portions may be sufficient for identification, they are not necessary. Results from the current experiments are consistent with these findings, in that performance in Meaning discrimination declined significantly when segment onsets and offsets were reduced in the vowels-only stimuli in Experiment 3—thereby likely also removing residual formant transitions.

2. Lateralization and temporal resolution

These observations also fit well with the view that lateralization effects in auditory processing primarily reflect temporal characteristics of the input (reviewed by Poeppel,

2003; Mirman *et al.*, 2004). Key elements of the argument are that functional features of sounds occur on varying time scales, that auditory processing includes multiple levels of temporal resolution, and that temporal resolution differs between the two cerebral hemispheres. Zatorre (1997) has for example argued that the left hemisphere is optimized for fine-grained temporal processing, while the right hemisphere is designed for coarse temporal processing that produces fine-grained spectral analysis. Poeppel (2003) suggests that both hemispheres actually perform both kinds of analyses, but with nonprimary auditory areas of the left hemisphere preferentially operating with shorter, 20–50-ms integration windows, and homologous regions of the right hemisphere preferring longer, 150–250-ms integration windows.

This temporal-dynamics approach to auditory processing makes no principled distinction between speech- and nonspeech sounds (Mirman *et al.*, 2004), or between linguistic and indexical cues within speech. Thus, in both Zatorre's and Poeppel's approaches, finer temporal processing is used when capturing rapidly changing aspects of sounds but compromises spectral analysis. These views suggest that the particular neural mechanisms recruited in speech processing can vary depending on the task involved. That is, left-hemisphere mechanisms should become primary when listeners are tracking temporally dynamic features such as consonantal contrasts and formant transitions in vowels. On the other hand, attending to more slowly changing features such as linguistic or affective prosody should bring longer temporal windows and right-hemisphere processing to the fore. The latter may not be critical to identifying phonetic units (e.g., Strange, 1989b), but may be crucial to hearing formant-based voice-quality variation related to talker-identity discrimination (e.g., Bachorowski and Owren, 1999). This view is also consistent with evidence that the neural systems involved in memory for word recognition and voices are functionally distinct, with verbal-memory tasks producing predominantly left-hemisphere activation and voice-memory tasks triggering right-lateralized activity (Stevens, 2004). It is thus important to investigate whether parallel differences exist among nonhuman mammals that also produce indexically rich vowel-like sounds, but without functional use of rapid temporal dynamics produced through articulation. Such studies could clarify whether differentiated processing of rapidly versus slowly changing auditory stimuli predated the emergence of linguistic vocalization, or instead was driven by selection advantages related to producing such signals.

D. Hypotheses based on current results

Results of the current experiments lead to several testable hypotheses. If the observations concerning coarticulation and phonemic restoration are correct, then meaning discrimination should be better if vowels and consonants are replaced with noise rather than being set to silence. However, that effect should be particularly prominent for vowels-only stimuli that retain more dynamic, co-articulation-related properties, such as those used in Experiment 2. The effect for stimuli with more stringently edited vowels, such as those used in Experiment 3, will likely not show more of an effect than consonants-only stimuli with noise fillers. In addition, meaning discrimination performance for words presented in sentence-context should be significantly enhanced relative to presentation in isolation. Here again the effect should be most significant for vowels-only stimuli, and less difference would be expected for consonants-only versions. In fact, if replacing with noise also produces forward and backward masking effects, replacing vowels with noise may have a deleterious effect on discrimination based on only consonants. Alternatively, if replacing with silence is creating misleading, linguistic-like cuing, then using noise should again improve performance. In contrast, editing using noise fillers is expected to have little effect on talker-discrimination results, whether testing at the word or sentence levels. In other words, testing should reveal more accurate talker identification from vowels-only than from consonants-only stimuli regardless of whether excised portions are replaced by silence or noise, and the length of the stimuli.

E. Conclusions

The outcomes from these experiments with human listeners may be particularly informative in a comparative/evolutionary perspective. Specifically, seeing important similarity between humans and nonhuman primates in production and perception of relatively steady-state, harmonically structured vowel(-like) signals, the same can be inferred for the early hominins that eventually gave rise to modern humans. The critical sound is one with a stable and relatively low-frequency fundamental that provides an inherently good "medium" for the individually distinctive imprint of the vocalizer's supralaryngeal vocal-tract filter. In this context, the current experiments underscore that in modern speech, this form of indexical cuing likely occurs as a vestige of a previous evolutionary condition. In other words, that listeners attending to the relatively stable components of vowel sounds for cues to vocalizer identity are relying on a speech component homologous with the individual calls produced by the hominin ancestors of modern humans. In attending to word meaning, however, listeners focus on the dynamic components that make this stream of continuously modulated sound a linguistic and uniquely human event.

ACKNOWLEDGMENTS

This work was partially supported by NIMH Prime Award 1 R01 MH65317-01A2, Subaward 8402-15235-X. We are grateful for helpful comments from two anonymous reviewers, one of whom pointed out the potential importance of whether the edited words began with a vowel or consonant. We thank Ronald A. Cole for making available unpublished material and associated stimuli, and Barbara Conboy for helpful comments on an earlier version of the manuscript. Thanks also to Michael Armer, Daniel Bauer, Jenna Burrell, Finn Christensen, Lucy Dunne, Suzanne Gifford, John Greek, Joshua Greenberg, Shaun Geer, Anna Herforth, Jennifer Hodge, Benita Hussain, Joanna Jackson, Malinda Lovic, Joseph Marasia, Lori Rosenberg, Samuel Scarano, and Nick Tapazoglou for recording the words used here. Fi-

nally thanks to members of the Psychology of Voice and Sound Laboratory for preliminary evaluation of word pairs.

Abercrombie, D. (**1967**). *Elements of General Phonetics* (Aldine, Chicago).
Ades, A. E. (**1977**). "Vowels, consonants, speech, and nonspeech," Psychol. Rev. **84**, 524–530.
Bachorowski, J.-A., and Owren, M. J. (**1999**). "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech," J. Acoust. Soc. Am. **2**, 1054–1063.
Bashford, Jr., J. A., Riener, K. R., and Warren, R. M. (**1992**). "Increasing the intelligibility of speech through multiple phonemic restorations," Percept. Psychophys. **51**, 211–217.
Bashford, Jr., J. A., and Warren, R. M. (**1987**). "Multiple phonemic restorations follow rules for auditory induction," Percept. Psychophys. **42**, 114–121.
Bauer, H. R. (**1986**). "A comparative study of common chimpanzee and human infant sounds," in *Current Perspectives in Primate Social Dynamics*, edited by D. M. Taub and F. A. King (Van Nostrand Reinhold, New York), pp. 327–345.
Berger, L. R. (**2001**). "Is it time to revise the system of scientific naming?" [retrieved from http://news.nationalgeographic.com/ news/2001/12/ 1204_hominin_id.html].
Bermajo, M., and Omedes, A. (**1999**). "Preliminary vocal repertoire and vocal communication of wild bonobos (*Pan paniscus*) at Lilungu (Democratic Republic of Congo)," Folia Primatol. **70**, 328–357.
Boatman, C., Hall, C., Goldstein, M. H., Lesser, R., and Gordon, B. (**1997**). "Neuroperceptual differences in consonant and vowel discrimination: As revealed by direct cortical electrical interference," Cortex **33**, 83–98.
Boersma, P., and Weenink, D. (**2005**). "Praat: Doing phonetics by computer," [computer program], retrieved from http://www.praat.org/.
Bond, Z. S., and Garnes, S. (**1980**). "Misperceptions of fluent speech," in *Perception and Production of Fluent Speech*, edited by R. Cole (Erlbaum, Hillsdale, NJ), pp. 115–132.
Boothroyd, A. (**1984**). "Auditory perception of speech contrasts by subjects with sensorineural hearing loss," J. Speech Hear. Res. **28**, 185–196.
Burkle, T. Z. (**2004**). "Contribution of consonant versus vowel information to sentence intelligibility by normal and hearing-impaired listeners," Masters thesis, Indiana University (unpublished).
Burkle, T. Z., Kewley-Port, D., Humes, L., and Lee, J. H. (**2004**). "Contribution of consonant versus vowel information to sentence intelligibility by normal and hearing-impaired listeners," J. Acoust. Soc. Am. **115**, 2601.
Caramazza, A., Chialant, D., Capasso, D., and Miceli, G. (**2000**). "Separable processing of consonants and vowels," Nature (London) **403**, 428–430.
Cole, R. (**1973**). "Different memory functions for consonants and vowels," Cognit Psychol. **4**, 39–54.
Cole, R., Yan, Y., Mak, B., Fanty, M., and Bailey, T. (**1996**). "The contribution of consonants versus vowels to word recognition in fluent speech," Proceedings of the ICASSP'96, pp. 853–856.
Cutting, J. E. (**1974**). "Two left-hemisphere mechanisms in speech perception," Percept. Psychophys. **16**, 601–612.
de Waal, F. B. M. (**1988**). "The communicative repertoire of captive bonobos (*Pan paniscus*), compared to that of chimpanzees," Behaviour **106**, 183–251.
Diesch, E., and Luce, T. (**1997**). "Magnetic mismatch fields elicited by vowels and consonants," Exp. Brain Res. **116**, 139–152.
Dubno, J. R., Dirks, D. D., and Langhofer, L. R. (**1982**). "Evaluation of hearing-impaired listeners using a nonsense-syllable test. II. Syllable recognition and consonant confusion patterns," J. Speech Hear. Res. **25**, 141–148.
Dudley, H. (**1939**). "Remaking speech," J. Acoust. Soc. Am. **11**, 169–177.
Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., and Tanenhaus, M. K. (**1995**). "Eye movements as a window into real-time spoken language comprehension in natural contexts," J. Psycholinguist. Res. **24**, 409–436.
Eimas, P. D. (**1963**). "The relation between identification and discrimination along speech and nonspeech continua," Lang Speech **6**, 206–217.
Fitch, T., Neubauer, J., and Herzel, H. (**2002**). "Calls out of chaos: The adaptive significance of nonlinear phenomena in mammalian vocal production," Anim. Behav. **63**, 407–418.
Fromkin, V. A. (1971). "The non-anomalous nature of anomalous utterances," Lang. **47**, 27–52.
Fu, Q.-J., and Shannon, R. J. (**1998**). "Effects of amplitude nonlinearity on phoneme recognition by cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **104**, 2570–2577.
Fujisaki, H., and Kawashima, T. (**1970**). "Some experimentation on speech perception and a model for the perceptual mechanism," Annual Report of the Engineering Research Institute (Faculty of Engineering, University of Tokyo), Vol. **29**, pp. 206–214.
Garner, W. R. (**1970**). "The stimulus in information processing," Am. Psychol. **25**, 350–358.
Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. (1990). "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM," National Institute of Standards and Technology, NTIS Order No. PB91-505065.
Goodall, J. (**1986**). *The Chimpanzees of Gombe* (Harvard University Press, Cambridge, MA).
Hardcastle, W. J., and Hewlett, N. (Eds.) (**1999**). *Coarticulation: Theory, Data, and Techniques* (Cambridge University Press, Cambridge).
Hauser, M. D., Evans, C. S., and Marler, P. (**1993**). "The role of articulation in the production of rhesus monkey (*Macaca mulatta*) vocalizations," Anim. Behav. **45**, 423–433.
Healey, A. F., and Repp, B. H. (**1982**). "Context independence and phonetic mediation in categorical perception," J. Exp. Psychol. Hum. Percept. Perform. **8**, 68–80.
Hillenbrand, J. M., and Nearey, T. M. (**1999**). "Identification of resynthesized /hVd/ utterances: Effects of formant contour," J. Acoust. Soc. Am. **105**, 3509–3523.
Hyde, J., and Linn, M. (**1988**). "Gender differences in verbal ability: A meta-analysis," Psychol. Bull. **104**, 53–69.
Jäncke, L., Wüstenberg, T., Scheich, H., and Heinze, H.-J. (**2002**). "Phonetic perception and the temporal cortex," Neuroimage **15**, 733–746.
Jenkins, J. J., Strange, W., and Miranda, S. (**1994**). "Vowel identification in mixed-speaker silent-center syllables," J. Acoust. Soc. Am. **95**, 1030–1043.
Kimura, D. (**1983**). "Sex differences in cerebral organization for speech and praxic functions," Can. J. Psychol. **37**, 19–35.
Lachs, L., McMichael, K., and Pisoni, D. B. (2000). "Speech perception and implicit memory: Evidence for detailed episodic encoding of phonetic events," Research on Spoken Language Processing Progress Report (Speech Research Laboratory, Indiana University), Vol. 24, pp. 149–167.
Ladefoged, P. (**2001**). *Vowels and Consonants: An Introduction to the Sounds of Languages* (Blackwell, New York).
Ladefoged, P., and Broadbent, D. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.
Ladefoged, P., and Maddieson, L. (**1996**). *The Sounds of the World's Languages* (Blackwell, Oxford).
Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., and Bourne, V. T. (**1976**). "Speaker sex identification from voiced, whispered, and filtered isolated vowels," J. Acoust. Soc. Am. **59**, 675–678.
Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (**1954**). "The role of consonant-vowel transitions in the perception of the stop and nasal consonants," Psychol. Monogr. **68**, 1–13.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psychol. Rev. **74**, 431–461.
Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (**1957**). "The discrimination of speech sounds within and across phoneme boundaries," J. Exp. Psychol. **54**, 358–368.
Lieberman, P. (**1975**). *On the Origins of Language: An Introduction to the Evolution of Human Speech* (Macmillan, New York).
Lindholm, J. M., Dorman, M., Taylor, B. E., and Hannley, M. (**1988**). "Stimulus factors influencing the identification of voiced stop consonants by normal-hearing and hearing-impaired adults," J. Acoust. Soc. Am. **83**, 1608–1614.
Macmillan, N. A., and Creelman, C. D. (**1991**). *Detection Theory: A User's Guide* (Cambridge University Press, Cambridge).
Macmillan, N. A., Goldberg, R. F., and Braida, L. D. (**1988**). "Resolution of speech sounds: Basic sensitivity and context memory on vowel and consonant continua," J. Acoust. Soc. Am. **84**, 1262–1280.
MacNeilage, P. F. (**1998**). "The frame/content theory of evolution of speech production," Behav. Brain Sci. **21**, 499–546.
MacNeilage, P. F., and Davis, B. (**2000**). "On the origin of internal structure of word forms," Science **288**, 527–531.
Marslen-Wilson, W. D. (**1987**). "Functional parallelism in spoken word-recognition," Cognition **25**, 71–102.
Miceli, G., Caltagirone, C., Capasso, R., Patria, F., Turriziani, P., and Caramazza, A. (**2004**). "The representation of segmental information: An fMRI investigation of the consonant-vowel distinction," Brain Lang **91**, 35–37.
Miller, J. L., and Eimas, P. D. (**1995**). "Speech perception: From signal to word," Annu. Rev. Psychol. **46**, 467–492.

Mirman, D., Holt, L. L., and McClelland, J. L. (**2004**). "Categorization and discrimination of non-speech sounds: Differences between steady-state and rapidly-changing acoustic cues," J. Acoust. Soc. Am. **116**, 1198–1207.

Monaghan, P., and Shillcock, R. C. (**2003**). "Connectionist modelling of the separable processing of consonants and vowels," Brain Lang **86**, 83–98.

Neely, J. H. (**1991**). "Semantic priming effects in visual word recognition: A selective review of current findings and theories," in *Basic Processes in Reading*, edited by D. Besner and G. W. Humphries (Erlbaum, Hillsdale, NJ), pp. 264–336.

Owren, M. J. (**2003**). "Vocal production and perception in nonhuman primates provide clues about early hominids and speech evolution," ATR Symp. HIS Series **1**, 1–19.

Owren, M. J., and Bachorowski, J.-A. (**2006**). "Hearing talker sex from brief vowel segments" (unpublished).

Owren, M. J., Seyfarth, R. M., and Cheney, D. L. (**1997**). "The acoustic features of vowels-like grunt calls in chacma baboons (*Papio cynocephalus ursinus*): Implications for production processes and functions," J. Acoust. Soc. Am. **101**, 2951–2963.

Parker, E. M., and Diehl, R. L. (**1984**). "Identifying vowels in CVC syllables: Effects of inserting silence and noise," Percept. Psychophys. **36**, 369–390.

Pisoni, D. B. (**1973**). "Auditory and phonetic memory codes in the discrimination of consonants and vowels," Percept. Psychophys. **13**, 253–260.

Pisoni, D. B. (**1975**). "Auditory short-term memory and vowel perception," Mem. Cognit. **3**, 7–18.

Poeppel, D. (**2003**). "The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'," Speech Commun. **41**, 245–255.

Rendall, D., Owren, M. J., and Rodman, P. S. (**1998**). "The role of vocal tract filtering in identity cueing in rhesus monkey (*Macaca mulatta*) vocalizations," J. Acoust. Soc. Am. **103**, 602–614.

Rendall, D., Owren, M. J., Weerts, E., and Hienz, R. D. (**2004**). "Sex differences in the acoustic structure of vowel-like grunt vocalizations in baboons and their perceptual discrimination by baboon listeners," J. Acoust. Soc. Am. **115**, 411–421.

Rendall, D., Rodman, P. S., and Emond, R. E. (**1996**). "Vocal recognition of individuals and kin in free-ranging rhesus monkeys," Anim. Behav. **51**, 1007–1015.

Rendall, D., Seyfarth, R. M., Cheney, D. L., and Owren, M. J. (**1999**). "The meaning and function of grunt variants in baboons," Anim. Behav. **57**, 583–592.

Repp, B. H. (**1984**). "Categorical perception: Issues, methods, findings, in *Speech and Language: Advances in Basic Research and Practice, Vol. 10*, edited by N. J. Lass (Academic, New York), pp. 243–335.

Repp, B. H. (**1992**). "Perceptual restoration of a missing speech sound-auditory induction or illusion?," Percept. Psychophys. **51**, 14–32.

Riede, T., Bronson, E., Hatzikirou, H., and Zuberbühler, K. (**2005**). "Vocal production mechanisms in a non-human primate: Morphological data and a model," J. Hum. Evol. **48**, 85–96.

Riede, T., and Zuberbühler, K. (**2003**). "The relationship between acoustic structure and semantic information in Diana monkey alarm vocalization," J. Acoust. Soc. Am. **114**, 1132–1142.

Samuel, A. G. (**1981**). "The role of bottom-up confirmation in the phonemic restoration illusion," J. Exp. Psychol. Hum. Percept. Perform. **7**, 1124–1131.

Shankweiler, D., and Studdert-Kennedy, M. (**1967**). "Identification of consonants and vowels presented to left and right ears," Q. J. Exp. Psychol. **19**, 59–63.

Shipley, C., Carterette, E. D., and Buchwald, J. S. (**1991**). "The effect of articulation on the acoustical structure of feline vocalization," J. Acoust. Soc. Am. **89**, 902–909.

Stevens, A. A. (**2004**). "Dissociating the cortical basis of memory for voices, words and tones," Cognit. Brain Res. **18**, 162–171.

Strange, W. (**1989a**). "Evolving theories of vowel perception," J. Acoust. Soc. Am. **85**, 2081–2087.

Strange, W. (**1989b**). "Dynamic specification of coarticulated vowels spoken in sentence context," J. Acoust. Soc. Am. **85**, 2135–2153.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (**1976**). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am. **60**, 213–224.

Su, T.-T. (**1999**). "Deletion phenomenon of phonemes in spontaneous Taiwanese-Mandarin," Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, pp. 379–382.

Teoh, S. W., Neuburger, H. S., and Svirsky, M. A. (**2003**). "Acoustic and electrical pattern analysis of consonant perceptual cues used by cochlear implant users," Audiol. Neuro-Otol. **8**, 269–285.

Titze, I. R., Baken, R. J., and Herzel, H. (**1993**). "Evidence of chaos in vocal fold vibration," in *Vocal Fold Physiology: Frontiers in Basic Science*, edited by I. R. Titze (Singular, San Diego), pp. 143–188.

Traunmüller, H. (**1994**). "Conventional, biological and environmental factors in speech communication: A modulation theory," Phonetica **51**, 170–183.

Välimaa, T. T., Määttä, T. K., Löppönen, H. J., and Sorri, M. J. (**2002**). "Phoneme recognition and confusions with multichannel cochlear implants: Consonants," J. Speech Lang. Hear. Res. **45**, 1055–1069.

Verbrugge, R. R., and Rakerd, B. (**1986**). "Evidence of talker-independent information for vowels," Lang Speech **29**, 39–57.

Warren, R. (**1999**). *Auditory Perception: A New Analysis and Synthesis* (Cambridge University Press, Cambridge).

West, P. (**1999**). "Perception of distributed coarticulatory properties of English /l/ and /r/," J. Phonetics **27**, 405–426.

Whiteside, S. P. (**1998a**). "Identification of a speaker's sex: A study of vowels," Percept. Mot. Skills **86**, 579–584.

Whiteside, S. P. (**1998b**). "Identification of a speaker's sex: A fricative study," Percept. Mot. Skills **86**, 585–591.

Wilden, I., Herzel, H., Peters, G., and Tembrock, G. (**1998**). "Subharmonics, biphonation, and deterministic chaos in mammalian vocalization," Bioacoustics 9, 171–196.

Wilson, M. D. (**1988**). "The MRC psycholinguistic database: Machine readable dictionary, version 2," Behav. Res. Methods Instrum. Comput. **20**, 6–11.

Zatorre, R. J. (**1997**). "Cerebral correlated of human auditory processing: Perception of speech and musical sounds," in *Acoustical Signal Processing in the Central Auditory System*, edited by J. Syka (Plenum, New York), pp. 453–468.

Zeng, F.-G., and Turner, C. W. (**1990**). "Recognition of voiceless fricatives by normal and hearing-impaired subjects," J. Speech Hear. Res. **33**, 440–449.

The use of visual cues in the perception of non-native consonant contrasts[a)]

Valerie Hazan,[b)] Anke Sennema,[c)] Andrew Faulkner, and Marta Ortega-Llebaria[d)]
Department of Phonetics and Linguistics, UCL, Gower Street, London WC1E 6BT, United Kingdom

Midori Iba
Konan University Institute for Language and Culture, 8-9-1 Okamoto, Higashi-nada-ku, Kobe-shi, Hyogo Japan

Hyunsong Chung[e)]
Department of English Education, Daegu University, 15, Naeri, Jinryang, Gyeongsan Gyeongbuk 712-714, South Korea

(Received 9 June 2005; revised 22 December 2005; accepted 22 December 2005)

This study assessed the extent to which second-language learners are sensitive to phonetic information contained in visual cues when identifying a non-native phonemic contrast. In experiment 1, Spanish and Japanese learners of English were tested on their perception of a labial/labiodental consonant contrast in audio (*A*), visual (*V*), and audio-visual (*AV*) modalities. Spanish students showed better performance overall, and much greater sensitivity to visual cues than Japanese students. Both learner groups achieved higher scores in the *AV* than in the *A* test condition, thus showing evidence of audio-visual benefit. Experiment 2 examined the perception of the less visually-salient /l/-/r/ contrast in Japanese and Korean learners of English. Korean learners obtained much higher scores in auditory and audio-visual conditions than in the visual condition, while Japanese learners generally performed poorly in both modalities. Neither group showed evidence of audio-visual benefit. These results show the impact of the language background of the learner and visual salience of the contrast on the use of visual cues for a non-native contrast. Significant correlations between scores in the auditory and visual conditions suggest that increasing auditory proficiency in identifying a non-native contrast is linked with an increasing proficiency in using visual cues to the contrast. [DOI: 10.1121/1.2166611]

PACS number(s): 43.71.Hw, 43.71.Es [ALF] Pages: 1740–1751

I. INTRODUCTION

The speech signal contains multiple acoustic cues to phonetic features, and this redundancy of information helps listeners in their native language to overcome degradation of the signal by noise, other voices, or reverberation. Speechreading is a further source of segmental information, as visible gestures can provide some cues to the place of articulation of vowels and the place and manner of articulation of consonants (e.g., Summerfield, 1983). The degree of information provided about these features is clearly more limited than that available auditorily. Indeed, the number of visemes in the language, i.e., the number of "visual categories" that are identifiable using speechreading alone, is markedly smaller than the number of phonemes (Walden *et al.*, 1977), and relatively poor consonant and vowel identification rates are achieved via speechreading alone (e.g., Binnie *et al.*, 1974). However, in many studies with listeners tested in their native language, the addition of speechreading cues has been shown to make speech more resistant to degradations due to noise and reverberation. For example, in an experiment in which listeners had to identify spondaic words in noise, adding visual cues was equivalent to improving the signal-to-noise ratio by 15 dB (Sumby and Pollack, 1954). More recent studies with more controlled experimental conditions (e.g., Nielsen, 2004) have shown that the performance advantage in noise may vary according to a number of factors, and that consonant contrasts vary in visual salience, but the general effect of visual cues shown in Sumby and Pollack's early study has been replicated many times. The term "*AV* benefit", defined as the amount of benefit derived by combining auditory and visual cues (Grant and Seitz, 1998), has been used to describe the advantage that is given by audio-visual presentation, especially for those listeners who cannot take full advantage of the auditory channel due to a hearing loss, for example. A common measure of *AV* benefit can be calculated from the difference between *AV* and *A* speech proportion-correct recognition scores expressed relative to the amount of improvement possible

[a)]Portions of this work were presented in "Audiovisual perception in L2 learners," Proceedings of ICSLP, Denver, September 2002, "The role of visual cues in L2 consonant perception, Proceedings of the 15th ICPhS, Barcelona, August 2003, and "Use of visual cues in the perception of a labial/labiodental contrast by Spanish-L1 and Japanese-L1 learners of English," Proceedings of ICSLP, Jeju, October 2004.

[b)]Electronic mail: val@phon.ucl.ac.uk
[c)]Current affiliation: Institute for Linguistics, University of Potsdam, Germany.
[d)]Current affiliation: Department of Spanish and Portuguese, University of Texas at Austin, Austin, TX.
[e)]Current affiliation: Department of English Education, Korea National University of Education, South Korea.

given the subject's *A* score (Sumby and Pollack, 1954). Grant and Seitz (1998) showed that *AV* benefit scores for listeners with hearing impairment substantially exceeded zero (mean of 0.672 for a consonant identification test), although the amount of benefit varied widely across listeners.

AV benefit will depend on the relative perceptual weighting of visual and auditory cues, and it does appear that this weighting varies across languages. Such weighting has been evaluated using the McGurk effect (McGurk and MacDonald, 1976), which is obtained when visual and acoustic cues to consonant identity are put in conflict (e.g., visual /g/ with auditory /b/ is typically perceived as /d/ if the two channels of information are closely integrated). This effect is much weaker for Japanese (Sekiyama and Tohkura, 1993) and Chinese listeners (Sekiyama, 1997) than for English or Spanish listeners (Massaro *et al.*, 1993). Visual cues are less informative in some languages, and this leads listeners to primarily direct their attention to auditory cues. For example, given that visual cues carry little information about intonation, auditory information will carry greater weight for tonal languages, as much contrastive information will come from the decoding of prosodic patterns. In consequence, providing voicing (F_0) information to deaf talkers of Cantonese, as a supplement to speechreading, leads to much greater improvement in speech recognition than providing the same information to a deaf talker of a nontonal language (Ching, 1986; Wei *et al.*, 1991). Characteristics of the phoneme inventory will also affect the relative weighting of visual and auditory information; indeed, languages which have fewer consonant contrasts based on visible place distinctions will be less visually marked than languages that are rich in place contrasts, especially where these are more forward in the vocal tract. Cultural reasons have also been advanced (Sekiyama and Tohkura, 1993) suggesting that auditory cues may become more heavily weighted in cultures in which it may be considered disrespectful to look the talker directly in the face.

That there are language-specific differences in the weighting of visual cues strongly suggests that the use of these cues is acquired in language development. Infants display sensitivity to visual information, as they prefer video presentations of talkers with congruent rather than incongruent auditory and visual components (Kuhl and Meltzoff, 1982; Patterson and Werker, 1999, 2003). However, there is also evidence that speechreading develops with age, as children aged 6 to 10 years are less influenced by visual information in their judgments of consonant identity than are adult learners (Massaro *et al.*, 1986, Sekiyama *et al.*, 2003). In a study looking at the development of visual influence in Japanese and English 6-years-old and adult listeners, it was shown that the level of visual influence was the same in 6 -years-old English and Japanese listeners, and that it increased significantly in English but not Japanese adults (Sekiyama *et al.*, 2003). This gradual development of speechreading mirrors increasing sensitivity to certain acoustic cues over the first 10 years of life (e.g., Nittrouer and Miller, 1997; Hazan and Barrett, 2000; Mayo and Turk, 2004).

Given language-specific differences in the informativeness and weighting of visual cues, it is important to consider whether learners make use of visual cues in a second language (L2) and whether they show evidence of clear *AV* benefit. It is known that the acquisition of a new language at the phonetic level (i.e., the perception and production of vowel and consonantal contrasts) can be effortful, especially in adulthood. This is because languages vary in their phoneme inventory, and children show much greater ability to discriminate native than non-native phonetic contrasts already by the end of the first year of life (Werker and Tees, 1984; Kuhl, 1993). The degree of difficulty of L2 acquisition is strongly linked to the relation between the phoneme inventories of the first and second language, as our knowledge and use of the phoneme inventory of L1 interferes with the acquisition of L2. Some models of L2 acquisition have made predictions for the degree of difficulty of acquiring different phonemic contrasts, but these models are based on knowledge gained from perceptual tests using auditory information alone. The Speech Learning Model (Flege, 1995) suggests that L2 sound categories that are phonetically similar to native categories tend to be assimilated to these categories by second-language learners. Categories that are phonetically dissimilar to native sound categories are seen as easier to acquire as there is no L1 interference, while categories that are identical in the two languages are typically perceived and produced without much difficulty. Best's Perceptual Assimilation Model (PAM: Best, 1995) is based on a different theoretical framework, that of the Direct Realist model of speech perception (e.g., Fowler, 1986), and it makes detailed predictions of L2/L1 assimilation on the basis of similarities in articulatory gestures between the L2 and L1 sounds. An L2 sound is either "categorized" as an exemplar of a native phoneme category, "uncategorized" if it is similar to two or more native categories, or "nonassimilable" if it is not at all similar to any native category. Both of these theories predict that certain phonemic contrasts will be difficult to acquire within a specific L1/L2 combination.

As noted above, current models of L2 acquisition primarily consider auditory cues to phonemic contrasts, and it is yet to be established how learners attune to visual distinctions marking phonemic contrasts that do not occur in the L1. The situation of L2 learners is quite different from that of native talkers who might be relying on visual cues because of a hearing loss, for example. Such talkers have knowledge of the phoneme categories of the language but lack full access to the auditory cues. L2 learners, on the other hand, might have quite good access to the auditory channel but will likely possess poor knowledge of specific phoneme categories in the L2 and therefore may have to learn to associate a particular visual cue with a phoneme label. As for auditory cues, we can examine the relation of L1 and L2 phonologies in terms of their viseme inventory and expect three different types of relations between the visual cues in the L1 and L2. The same visual cue may occur in both languages to mark the same phoneme distinctions. We would expect in this case that the L2 distinction will be assimilated to the L1. Alternatively, a visual cue in an L2 may have no counterpart in the L1. An example of this arises for labiodentals, which occur

in English but not in Japanese. Finally, a visual cue may occur in the L1 but for different phoneme distinctions than in the L2. For example, the labial/labiodental distinction contrasts /b-p/ from /f-v/ in English, but in Spanish only contrasts /b-p/ from /f/ as there is no voiced labiodental fricative. An aim of our study is to see whether experience of a particular visual cue in an L1 will aid the acquisition of a novel phoneme contrast in L2.

Previous studies of visual cues for phoneme identification in an L2 present a mixed picture in terms of the benefit they bring the L2 learner. One study of L2 learners' use of visual information investigated a wide range of English consonants and vowels presented in mild background noise in auditory and audio-visual conditions by Spanish-L1 listeners and native controls (Ortega-Llebaria *et al.*, 2001). Consonant confusions that were language dependent, and which could be attributed to differences in the phoneme inventories of Spanish and English, were not reduced by the addition of visual cues, whereas confusions that were common to both listener groups and were predictable from acoustic-phonetic similarities between categories were reduced in the audio-visual condition. This suggested that the relation between L1 and L2 phoneme inventories affected the use of visual cues in an L2. A wider-ranging study of congruent and discongruent audio-visual syllables with listeners from four different L1 backgrounds showed that the influence of the visual cue was "dependent upon its information value, the intelligibility of the auditory cue and the assessment of similarity between the two cues" and was also affected by linguistic experience (Hardison, 1999). A study investigating the effect of visual cues on the perception of spoken words excised from scripted and unscripted sentences using a gating method did show faster recognition of audio-visual stimuli by non-native (Japanese) speakers, but complex interactions with talker-related and content-related effects limit the interpretability of this result (Hardison, 2005). There is some evidence that listeners who do not weigh visual cues heavily in their native language (e.g., Japanese speakers) attend more to visual cues (as shown by a larger McGurk effect) when listening in a second language, but this evidence is based on a relatively small amount of data (Sekiyama *et al.*, 1996).

The aim of our study was to evaluate L2 learners' sensitivity to visual cues for English phonemic contrasts that do not occur in their native phoneme inventory. In order to evaluate the effect of visual salience, experiments evaluated the perception of two English phonemic contrasts differing in the visual distinctiveness of their articulatory gestures: the highly distinctive contrast between labial (/b/-/p/) and labiodental (/v/) consonants and the less visually distinctive contrast between /r/ and /l/. Each of these two contrasts was tested with learners of English from two different language backgrounds in order to evaluate the effect of the presence or absence of the visual cues in the native phoneme inventory on the acquisition of the non-native contrasts.

Our first study investigated the perception of the labial-labiodental contrast by using a three-way contrast between the English sounds /v/, /b/, and /p/. The contrasts between /b/ and /v/ and between /b/ and /p/ are both difficult for Japanese and for Spanish learners of English. For the Spanish-L1 group, the following L1-linked confusions were predicted given the differences between the consonantal systems in Spanish and English (e.g., O'Connor, 1967). First, the English voiced plosive /b/ was expected to be confused with voiceless plosive /p/ as both English /b/ and Spanish /p/ are phonetically realized as unaspirated plosives and have similar voice onset times (Abramson and Lisker, 1972; Ortega-Llebaria *et al.*, 2001). In Spanish, /b/ is realized as the allophone [b] after a pause or a nasal consonant, and as the allophone [β] elsewhere, i.e., [be$\beta\acute{e}$] "bebé" *baby*, [um be$\beta\acute{e}$] "un bebé" *a baby*; [la βeβé] "la bebé" *the baby* (Navarro-Tomás, 1977; Quilis, 1993). Spanish does have an unvoiced labiodental /f/ in its phoneme inventory, which is realized as [v] when it is in coda position and followed by a voiced consonant, i.e., /afganistán/ [avɣanistán]. However, the incidence of [v] in Spanish is very limited since only a few loan words have /f/ in coda (Hualde, 2005). Therefore, it would be expected that listeners would be sensitive to the difference in visual cues between labials and labiodentals as both occur in Spanish. In Japanese, there is a voicing contrast between /p/ and /b/ but /p/ tends to be unaspirated so that there is potential assimilation of the English /b/ to the Japanese /p/. The English /v/ does not occur in the Japanese consonantal system but the Japanese /b/ is sometimes realized as the bilabial approximant [β] when between vowels so that confusions of English /v/ and /b/ could be expected. There are no labiodentals in Japanese so Japanese listeners would not be exposed to a visual contrast between labials and labiodentals. Our research hypothesis was that Spanish listeners would make better use of visual cues in their perception of the contrast in English between /v/ and /b-p/ because they would transfer their knowledge of the labiodental gesture (/f/) from their L1, even if /v/ does not occur in Spanish. In contrast, Japanese-L1 listeners will be less sensitive to this visual distinction because it does not occur in their L1.

The second study investigated the perception of the contrast between the English phonemes /l/ and /r/. This is a less visually salient contrast than the labial/labiodental one, even for native speakers. This contrast has been extensively used in studies of L2 speech perception as it is difficult for Japanese and Korean learners of English to acquire due to the different phonological status of this contrast in these languages. In Japanese, there is a single liquid within the phoneme inventory. Although phonetic variation in the production of Japanese /r/ occurs across talkers, this phoneme is most often described as an alveolar flap [ɾ], and it occurs at syllable onset only, in word-initial or word-medial position. The English phonemes /l/ and /r/ both tend to be assimilated, with variations in the goodness of fit, to the native alveolar flap, leading to problems in the discrimination and identification of these English phonemes (Bradlow *et al.*, 1997, Iverson *et al.*, 2003, Aoyama *et al.*, 2004). The /l/-/r/ contrast is particularly difficult for Japanese learners because it is primarily marked by differences in third formant transitions to which Japanese listeners are not particularly sensitive, as they tend to give greater weight to more variable secondary cues such as *F*2 transition (Yamada, 1995; Iverson *et al.*, 2003; Gordon *et al.*, 2001; Lotto *et al.*, 2004).

In modern standard South Korean, a liquid is phonetically realized as an [l], [ɾ], and [n]. In intervocalic positions, a liquid becomes a flap [ɾ], or becomes a lateral [l] in intervocalic geminates. There is therefore the expectation that the English /l/-/r/ contrast might be easier for Korean-L1 learners than for Japanese-L1 learners in intervocalic position, but that in initial position, Korean-L1 learners could have greater perceptual difficulties. These effects were indeed found in a study of the perception of the Australian English /l/-/r/ contrast by Japanese-L1 and Korean-L1 learners (Ingram and Park, 1998). Because Korean-L1 learners will have at least some experience of visual cues to the lateral gesture in some syllable positions whereas Japanese-L1 will not, we therefore expect that Korean-L1 learners will make better use of the visual information to the contrast than their Japanese counterparts.

In summary, our hypothesis was that greater evidence of *AV* benefit would be shown for the labial/labiodental contrast because of its greater visual salience, but also that the degree of *AV* benefit would vary across L2 groups and that it would be strongly influenced by the visual cue inventory in the learners' L1. For the labial/labiodental contrast, we hypothesized that greater *AV* benefit would be seen for Spanish-L1 listeners, as these listeners are exposed in their first language to a visual contrast between labials and labiodentals, even if the /b-p/-/v/ contrast itself is absent. For the /l/-/r/ contrast, we hypothesized better use of the visual cues by the Korean-L1 learners because they have had some exposure to laterals in Korean whereas Japanese-L1 learners have not.

II. EXPERIMENT 1: LABIAL/LABIODENTAL CONTRAST

A. Speech material

Nonsense words were used in this experiment as consonant identification in an L2 is significantly affected by lexical familiarity (e.g., Aoyama *et al.*, 2004). The consonants /p/, /b/, and /v/ were embedded within nonsense words with the following structure: CV, VCV, or VC, where V was one of the following: /i,ɑ,u/ (see the full list of words in Appendix A).

B. Talker and recording procedures

A phonetically trained female talker of South Eastern British English recorded the test items. Video recordings were made in a soundproof room, with the talker's face set against a blue background and illuminated with a key and a fill light. The talker's head was fully visible within the frame. Video recordings were made to a Canon XL-1 DV camcorder. Audio was recorded from a Bruel & Kjaer type 4165 microphone to both the camcorder and to a DAT recorder. The video channel was digitally transferred to a PC and time aligned with the DAT audio recording, which was of higher quality than the audio track of the video. Video clips were edited so that the start and end frames of each token showed a neutral facial expression. Stimuli were down-sampled once the editing had been completed (250×300 pixels, 25 f/s, audio sampling rate 22.05 kHz).

C. Listeners

Experiment 1 involved 32 Spanish learners of English, 47 Japanese learners of English, and 12 native controls. Thirty-six Spanish listeners participated in the experiment in total, but four obtained scores of 100% in the *A* condition and were consequently excluded from further analyses. Of the remaining 32 listeners, 28 were university students of the Universitat Autónoma de Barcelona tested in Spain and four were attending an intensive language course at a School of English in London. They were chosen on the basis that they were at a lower- to lower-intermediate level of English proficiency, and were monolingual native talkers of Castilian Spanish. The listeners were aged between 18 and 23 years (mean 19.3 years) and they had been learning English for between 2 and 12 years (mean duration of English instruction 6.9 years), but none had any special English language training. Of the 32 subjects, 18 had never been to an English-speaking country, and 12 had spent between 1 to 4 months in an English-speaking country. Only two of the listeners had lived in an English-speaking country for more than 6 months. All participants reported normal hearing and normal or corrected vision. The listeners were volunteers and received a small payment for their participation in the test.

Forty-seven Japanese listeners participated in the experiment. All were students of the English Department of Konan University and were tested in Japan. They were from a variety of degree programs and were taking an English elective course. They were aged between 19 and 23 years, had already learned English for an average of 6 years at high school before entering university, but did not often practice listening to English sounds.

A group of 12 native speakers of British English (5 female, 7 male), who were unfamiliar with the talker, served as control group and identified the test items in the two blocks of the video-alone condition. They ranged in age from 21 to 40 years (median 26 years, mean 27.8 years). Of these 12 controls, 9 were students recruited from University College London who reported having experience with L2 learners of English and three were trained EFL teachers.

D. Experimental task

A program for the computer-based presentation of the closed-set identification task was designed using the CSLU toolkit software (Cole, 1999). In order to make the test program more appealing to the subjects, a talking head ("Baldi," as described in Massaro, 1998) was used on-screen to explain the task to the listener and to give general feedback on the percentage of correct responses at the end of each section of the test. It must be noted that the talking head was only used to give instructions and that all audio-visual stimuli used in the perception tests were natural. English subtitles were always included with the talking head. An experimenter also gave instructions at the beginning of the first test session to ensure that learners comprehended the task. All experimental instructions were given in English. Learners were told that they would hear a meaningless word and that they had to decide which consonant they had heard in the word.

There were three test conditions: (1) Video alone (*V*); (2) Audio alone (*A*); and (3) Audio-visual (*AV*), with two blocks of 81 items per condition (three tokens for each consonant and vowel in three positions). Fifty-four repetitions of each consonant (across vowels and positions) were therefore presented to each listener in each test condition. The test took about 40–45 min to complete.

Before each section of the test, listeners received specific instructions from the talking head about which condition would be following next and what they had to attend to. In the *A* condition, subjects were told to attend to the speech sound, in the *V* presentation to look at the screen and to attend to lips of the talker for possible cues. In the *AV* condition they were directed to make use of both audio and video information. After the instruction, the test started with no further practice session. Subjects entered their response after the stimulus was played by clicking on a three-button box on-screen showing the three test consonants. Each stimulus was presented only once with no possibility of repetition. After the subjects had entered their response, the program continued automatically with the next stimulus. No feedback was given for individual responses but subjects were given an indication of the mean percentage of correct responses per test consonant and of the overall mean percentage of correct identification at the end of each section of the test.

The order of items was randomized within each block for each listener. Two orders were used for the presentation of the three conditions, *AV*, *A*, *V* or *A*, *AV*, *V*, and the two orders were counterbalanced across listeners. The *V* condition was presented last in both orders as it was likely to be the most difficult condition for the learners. As each block only took around 6 min to complete, it is unlikely that results for the *V* condition would have been depressed due to task boredom or fatigue.

In all locations, the test was run on PCs in quiet surroundings. The video files appeared in the center of the computer screen in a window of 12×9 cm (varying slightly with screen size). Audio stimuli were presented to both ears at a comfortable listening level via headphones. As the experiment was run in three different countries, headphone specifications may have varied across groups. However, our experimental tasks simply required good sound reproduction in the range from about 500–3000 Hz and previous tests of many different types of headphones suggest that all are adequate in that range (see also Iverson *et al.*, 2005).

The 12 British English control listeners identified 81 items in each of the two blocks of the video-alone presentation (*V*). They followed the same test methodology (if reduced to one condition) and received the same instructions as the Japanese and Spanish listeners.

E. Results

1. Group results

Although the /b/-/p/ contrast had been included in this study because confusions were expected between these two English consonants, we were principally interested in whether the listeners were able to perceive the distinction which is marked visually, i.e., the difference between the bilabial and labiodental consonants. Voicing confusions (/b/ perceived as /p/ and /p/ perceived as /b/) were therefore ignored and analyses focused on the perception of the contrast in place of articulation. One way of looking at the correct detection of the labiodental feature is simply to look at correct /v/ identification. For Spanish-L1 speakers, mean percentages were 82.3% (s.d. 22.2) in the *V* condition, 87% (s.d. 12.3) in the *A* condition, and 88.6% (s.d. 14.4) in the *AV* condition. For Japanese-L1 speakers, mean percentages were 50.7% (s.d. 18.8) in the *V* condition, 64.1% (s.d. 19.2) in the *A* condition, and 65.5% (s.d. 19.4) in the *AV* condition. Given that there were three possible response labels (*P*, *B*, and *V*), it is important to correct for any potential bias in the responses given by the learners. The unbiased signal detectability measure, *d* prime (d'), was used for this purpose (Green and Swets, 1966). It was calculated as the difference between the standardized *z* scores for hit-rate ("*V*" responses to /v/) minus that of the false-alarm-rate ("*V*" responses to /p/ and /b/). In order to avoid infinite *z* scores, hit rates of 1.0 and false-alarm rates of 0 were adjusted to correspond to 0.5 errors. Despite there being twice as many labial as labiodental stimuli, there was no evidence of bias as shown by average beta values close to 1.0 in all conditions. Mean scores in each condition are displayed in Fig. 1. Analyses of variance (ANOVAs) for repeated measures were then applied to the d' scores to look at the within-subject effect of test condition (*A*, *AV*, and *V*) and at the between-subjects effect of L1 background (Spanish-L1 or Japanese-L1).

The effect of condition was statistically significant $[F(2,154)=19.05;\ p<0.001]$. Pairwise comparisons with Bonferroni adjustments show that d' for place perception was significantly lower in the *V* condition than in the *A* and *AV* conditions, and that it was significantly better in the *AV* than in the *A* condition. The effect of L1 background was also significant $[F(1,77)=53.01;\ p<0.0001]$ with Spanish-L1 learners achieving higher scores than Japanese-L1 learners. The condition by L1 interaction was not significant.

Performance in the visual condition was further examined for differences in the use of visual information across L1 groups. Visual place perception showed a d' score of 4.61 (s.d. 1.38) for the native controls, as compared to 3.10 (s.d. 1.11) for the Spanish-L1 learners and 1.49 (s.d. 1.03) for the Japanese-L1 learners. One-way analyses of variance on scores in the visual condition showed performance for native controls to be significantly better than performance for Spanish-L1 learners, which was in turn significantly better than that for Japanese-L1 learners.

2. Individual results

In the left-hand panel of Fig. 2, a crossplot of place perception scores in the *A* and *V* conditions is presented for individual Japanese-L1 and Spanish-L1 listeners. High-performing individuals (mostly from Spanish-L1 background) with place perception at ceiling level in the *A* condition also had near-perfect place perception using visual cues. The level of performance in Japanese-L1 listeners was

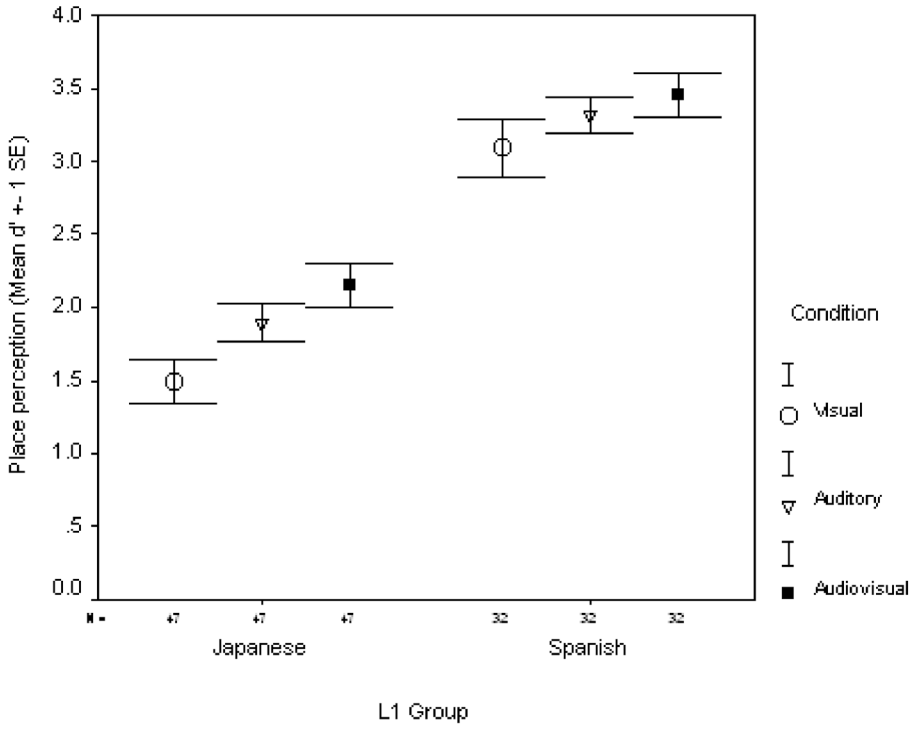


FIG. 1. Detectability (d') of the labial/labiodental distinction (i.e., place perception) in the audio (A), visual (V), and audio-visual (AV) conditions for the Japanese-L1 and Spanish-L1 learners. The error bars show 1 standard error from the mean. As reference, the mean detectability score (d') for native controls in the V condition was 4.61.

typically lower but there is evidence of a subgroup with d' in excess of 1 with auditory cues but random performance (d' near to 0) with visual cues.

The data were then examined to look at the degree of AV benefit seen for individual learners. The criterion for individual AV benefit used here was that d' for the AV condition exceeded that in the A condition by at least 1 d' unit. (i.e., by 1 standard deviation). The left-hand panel of Fig. 3 plots individual d' scores for the AV condition as a function of the A d' score. The lines represent a bound of $+/-$ 1 d' unit around a slope of 1 [i.e., $d'(AV)=d'(A)$]. Only 8 out of the 83 learners (9.6%) are outside these bounds. Since this proportion is not more than the 15% that would be expected by chance, the data do not clearly signal AV benefit for individual subjects, notwithstanding the overall AV advantage for the group shown in ANOVA.

F. Discussion

In summary, there is evidence of AV benefit in the perception of the labial/labiodental place of articulation for both learner groups and no indication of substantial individual variability in subjects' ability to improve their perception in the audio-visual condition relative to auditory presentation alone. Generally, better place perception was obtained for Spanish than for Japanese listeners using auditory cues. The

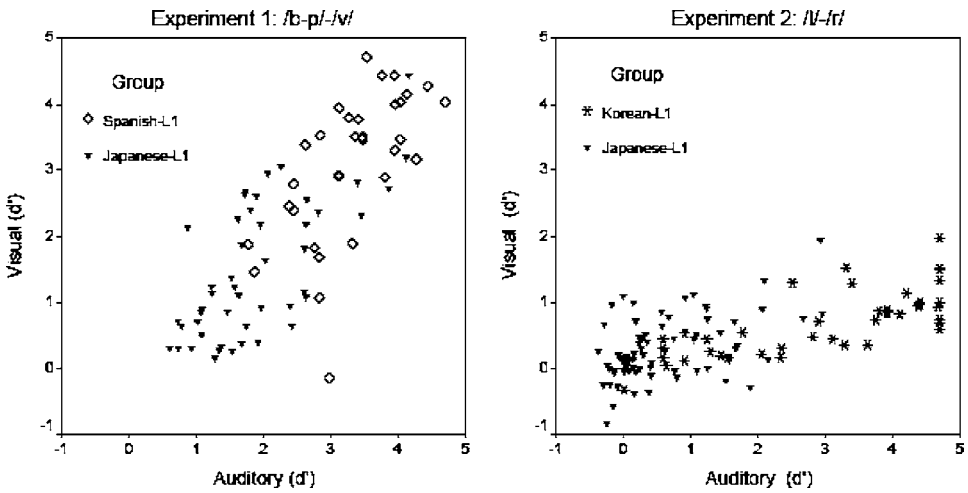


FIG. 2. Crossplot showing individual scores in the auditory (A) and visual (V) conditions in the two experiments. In the left-hand panel, place perception scores (d') for the /b-p/-/v/ contrast are plotted for individual Japanese and Spanish learners. In the right-hand panel, /l/-/r/ identification scores (d') are plotted for individual Japanese and Korean learners.

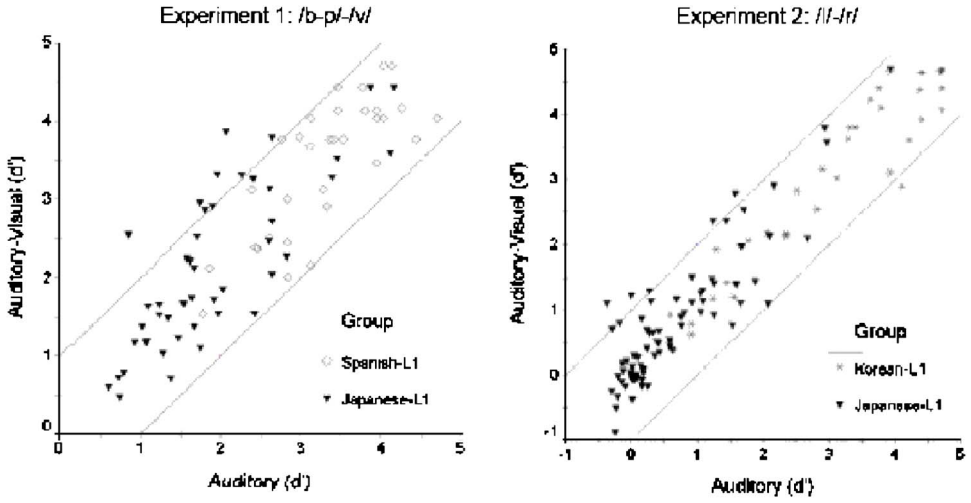


FIG. 3. Crossplot showing individual scores in the auditory (A) and audio-visual (AV) conditions in the two experiments. The two lines represent a bound of $+/-$ 1 d' unit around a slope of 1 [i.e., $d'(AV)=d'(A)$]. In the left-hand panel, place perception scores (d') for the /b-p/-/v/ contrast are plotted for individual Japanese and Spanish learners. In the right-hand panel, /l/-/r/ identification scores (d') are plotted for individual Japanese and Korean learners.

explanation for this may be linked to the specific language exposure of our learners tested in Barcelona. First, these learners are exposed to [v] as allophonic realizations of /f/ in coda in Spanish and in Catalan (Prieto, 2004). Further, the Spanish group may have been exposed to voicing contrasts in onset position, in Catalan, for fricatives at other places of articulation. The exposure of our Japanese-L1 learners to languages which have a /v/ phoneme is likely to be much more limited.

There was also a distinct difference in the "visual-alone" performance across L1 groups. For the Spanish-L1 group, on average, place perception was not much different in the visual-alone than in the audio-alone condition; for the Japanese-L1 group, auditory-alone perception was significantly better than the visual-alone perception, which suggests better use of acoustic than visual cues. As discussed previously, Spanish-L1 listeners are exposed to the visual distinction between labials and labiodentals in their L1 (e.g., /b/-/f/ contrast), and they seem able to associate the labiodental gesture with the voiced counterpart of the native labiodental /f/. A more general difference between the Japanese-L1 and Spanish-L1 learners is in the relative weighting of auditory and visual cues in their native language. Spanish-L1 speakers show a much stronger McGurk effect than Japanese-L1 speakers, despite having similar phonetic inventories (Massaro *et al.*, 1993). Both these factors could explain the greater ease in acquiring both the auditory and visual cues to the labial/labiodental distinction for Spanish-L1 learners of English.

It must be noted that there is variability in the use of visual cues within the subgroup of Japanese-L1 learners who are displaying good place perception in the auditory condition: some also show relatively high scores in the visual-alone condition and therefore have learned to associate the labiodental gesture with the appropriate phoneme label, while others seem little able to use visual information to the contrast. Therefore, although the acquisition of the labial/labiodental contrast via exposure to the L2 does generally seem to increase sensitivity to visual cues to the contrast, it is by no means a universal process across all learners.

III. EXPERIMENT 2: /l/-/r/ CONTRAST

A. Speech material

The two English consonants /l/ and /r/ were embedded in initial and medial position in nonsense words in the context of the vowels /i,ɑ,u/. Because different predictions are made in terms of the perception of /l/-/r/ in clusters according to L1 background, both singletons and clusters were included: the consonants were presented as singleton or cluster with the initial consonant of the cluster being /k/ and /f/ and appeared in the structure CV, cCV, VCV, and VcCV (see the full list of words in Appendix B).

B. Talker and recording procedure

A phonetically trained female talker of South Eastern British English recorded the test items. This was a different talker from the one recorded in experiment 1. The same recording procedure was used as in experiment 1. Three items were produced for each consonant in each syllabic and vowel context (27 initial /l/ and /r/, 27 medial /l/ and /r/), yielding a total of 108 items.

C. Listeners

Experiment 2 involved 78 Japanese learners of English and 42 Korean learners of English. Of the Japanese-L1 listeners, 42 were university students of Kochi University and tested in Japan, 20 students were attending a summer course in phonetics at UCL, 9 were recruited from a School of English in London, and 7 were students of a preacademic language course at UCL. The median age was 20.0 years (mean: 20.3 years) with an age range from 17 to 32 years; the median number of years of English study was 8 years (mean: 8.0 years) with a range from 4 to 20 years. The learners were at a lower- to lower-intermediate level of English proficiency and their previous English study had emphasized grammar and reading rather than listening to English sounds. Only three of the students had spent more than 6 months and three students 1–5 months in an English-speaking country; the remaining 72 students had never lived in an English-speaking country.

All of the Korean-L1 listeners were university students of Daegu University and tested in Korea. The learners were at a lower- to lower-intermediate level of English proficiency. The median number of years of English study was 10 years (range: 6 to 18 years) but with little focus on spoken English. Six of the students had spent more than 6 months in an English-speaking country. All participants reported normal hearing, and normal or corrected vision.

The same group of 12 native speakers of British English as in experiment 1 served as the control group in experiment 2. They followed the same experimental protocol as the non-native subjects, but were only tested in the *V* condition. Each listener in the control group identified 108 presentations of each consonant (across vowels and positions).

D. Experimental task

The same experimental procedure and testing software were used as in experiment 1. The instructions given were adapted from experiment 1 with the difference that in the *V* condition subjects were asked to look at the screen and to attend to lips and, if visible, tongue movement. The two test consonants were shown on a two-button display on the computer screen and subjects were asked to indicate their choice by clicking on the appropriate button with a mouse. After these instructions, subjects started the test with no further practice session. Each token was played once and after the subjects had given their response the program continued automatically to the next token. No feedback was given on whether individual responses were correct or incorrect but learners were given an indication of the mean percentage of correct responses per test consonant and of the overall mean percentage of correct identification at the end of each section of the test. It took about 50 min to complete the test (in the *A*, *V*, and *AV* conditions).

The items were presented in three conditions: audio-alone (*A*), visual-alone (*V*), and audio-visual presentation (*AV*). Two blocks of 108 items were presented per condition and each block took about 8 min to complete. Each listener responded to 108 repetitions of each consonant (across vowels and positions) in each test condition. The order of items was randomized within each block. Two orders were used for the presentation of the three conditions: *AV*, *A*, *V* or *A*, *AV*, *V*, and the two orders were counterbalanced across listeners. As in experiment 1, the *V* condition was always presented last as it was likely to be the most difficult condition for the learners.

E. Results

1. Group results

The percentage of correct /l/ and /r/ responses was calculated (see Fig. 4). For Japanese-L1 learners, mean /l/-/r/ identification was 55.8% (s.d. 8.7) for the *V* condition, 61.5% (s.d. 13.7) for the *A* condition, and 63.4% (s.d. 14.8) for the *AV* condition. For the Korean-L1 learners, mean /l/-/r/ was 63.1% (s.d. 8.7) for the *V* condition, 88.3% (s.d. 14.3) for the *A* condition, and 87.9% (s.d. 14.6) for the *AV* condition. As in experiment 1, scores were then converted to the signal detectability measure *d* prime (*d′*). For the Japanese group,

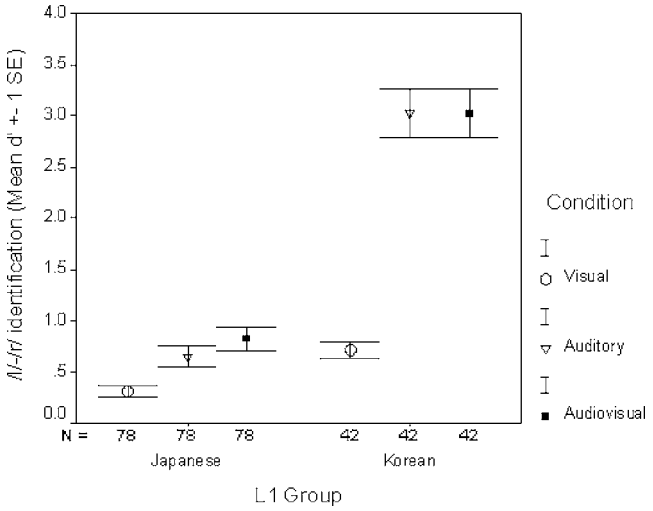


FIG. 4. /l/-/r/ identification (*d′*) in the audio (*A*), visual (*V*), and audio-visual (*AV*) conditions for the Japanese-L1 and Korean-L1 learners. The error bars show 1 standard error from the mean. As reference, the mean detectability score (*d′*) for native controls in the *V* condition was 1.31.

because some learners were tested in the UK and others in their home country, a repeated-measures ANOVA was run to see whether the place of testing affected performance. As this effect was not significant, the data were grouped per language background for the main analyses. The native controls were not tested on the *A* and *AV* conditions because of likely ceiling effects but achieved a mean score of 72.1% (*d′* of 1.31) on the *V* condition. Therefore, for native listeners, identification using visual cues alone is above chance, although it appears that the contrast is not highly salient visually when compared to a *d′* of 4.61 obtained for the same subjects in the *V* condition for the labial/labiodental contrast in experiment 1.

A repeated-measures analysis of variance evaluated the within-subject effect of condition (*V*,*A*,*AV*) and the between-subject effect of L1 background. As suggested by the mean scores, the effect of L1 background was significant with Korean-L1 learners achieving higher scores than Japanese-L1 learners [$F(1,118)=95.40$; $p<0.001$]. The effect of test condition was significant [$F(2,236)=195.88$; $p<0.001$]. Pairwise comparisons with Bonferroni adjustments showed that this was due to lower scores being obtained in the *V* condition than in the *A* and *AV* conditions, but there was no significant difference between the *A* and *AV* conditions. The condition by L1 background interaction was also significant [$F(2,236)=93.48$; $p<0.001$]. Examination of means suggests that this is due to scores for *A* and *AV* (but not *V*) conditions being higher for Korean than Japanese listeners.

To check that the difference in performance across L1 groups was not due to significant differences in amount of English instruction, statistics were rerun on data from a subgroup of listeners with a more narrow range of years of learning (7 or 8 years of exposure to English). This subgroup included 41 Japanese-L1 and 10 Korean-L1 speakers. The same effects were obtained for this subgroup as described above for the complete dataset.

To check that ceiling effects were not masking the effect

TABLE I. Mean /l/-/r/ identification scores (d') for word-initial and word-medial singletons and clusters obtained in the three conditions (V,A,AV) by Japanese-L1 and Korean-L1 learners of English in experiment 2. Standard deviations are given in parentheses.

	Japanese-LI			Korean-LI		
	V	*A*	*AV*	*V*	*A*	*AV*
Initial singleton	0.66 (0.74)	1.05 (1.08)	1.31 (1.23)	1.19 (0.80)	2.65 (1.21)	2.71 (1.18)
Initial cluster	0.25 (0.51)	0.65 (1.01)	0.74 (1.10)	0.66 (0.57)	3.18 (1.46)	3.16 (1.51)
Medial singleton	0.55 (0.79)	0.99 (1.10)	1.16 (1.20)	1.06 (0.57)	2.58 (1.19)	2.69 (1.26)
Medial cluster	0.25 (0.46)	0.63 (1.02)	0.75 (1.08)	0.54 (0.51)	3.11 (1.57)	2.97 (1.54)

of *AV* benefit (in Korean-L1 learners, especially), statistics were run on the data for the lowest 50% of listeners when ranked on their mean performance across the three test conditions for each L1 group. Mean d' for the two subgroups were as follows: Japanese-L1 group *A* 0.46, *AV* 0.94, *V* 0.49; Korean-L1 group *A* 1.76, *AV* 1.72, and *V* 0.36. Repeated-measures ANOVAs were run on this selected group. The effect of condition was significant [$F(2,116)=61.468$; $p<0.001$] and Bonferroni-adjusted comparisons showed that this was due only to mean performance in the *V* condition being significantly lower than for *A* and *AV* conditions. The condition by L1 group interaction was also significant [$F(2,116)=57.91$; $p<0.001$] which seems to be due to the higher identification rate for Korean-L1 listeners in the *A* and *AV* but not the *V* condition. Therefore, even for this lower-performing group, there was no evidence of *AV* benefit for either group of listeners.

The effects of consonant, consonant position, and syllable structure were examined in more detail, given the difference between the likely patterns of assimilation of L2 consonants to L1 categories in the two L1 groups. Identification scores were calculated for each learner for the following four categories in each condition: singletons in word-initial position, singletons in word-medial position, clusters in initial position, and clusters in word-medial position. Based on phonological differences between Japanese and Korean, Ingram and Park (1998) predicted different patterns of difficulty for Korean and Japanese learners of English acquiring the /l/-/r/ contrast. For Korean-L1 learners, the following order of difficulty was expected (from most difficult to easiest): word-initial singletons, word-initial clusters, word-medial singletons. Japanese-L1 learners were expected to have greatest difficulty with word-initial clusters and to find word-initial and word-medial singletons equally difficult to identify. Ingram and Park do not include predictions for word-medial clusters.

ANOVAs were carried out to evaluate the between-subject effect of L1 group and within-subject effect of syllable structure. In the auditory condition, none of the main effects were significant but there was a significant interaction between syllable structure and L1 group [$F(1,118)=47.35$; $p<0.001$]. Observation of the data suggests that lower scores were obtained for clusters than for singletons for Japanese-L1 learners, as expected, and that lower scores were obtained for singletons than for clusters for Korean-L1 learners (See Table I). Results therefore generally support the predictions of Ingram and Park (1998).

In the visual-alone condition, the following effects were obtained. The main effect of syllable structure was significant with lower scores obtained for /l/-/r/ in clusters than in singletons [$F(1,118)=105.77$; $p<0.001$]. The main effect of position was also significant, with lower scores obtained in word-medial than in word-initial consonants [$F(1,118)=5.79$; $p<0.02$]. The interaction between syllable structure and L1 group was also significant [$F(1,118)=4.05$; $p<0.02$]. This seems to be due to better discriminability for singletons for Korean-L1 learners. The patterns of difficulty were therefore different from those seen in the auditory condition and seemed to be determined by the visual salience of the contrast in different consonant structures and positions.

2. Individual results

In the right-hand panel of Fig. 2, a crossplot of /l/-/r/ perception in the *A* and *V* conditions is presented for individual Japanese-L1 and Korean-L1 listeners. The level of performance in Japanese-L1 listeners was typically lower than that of Korean-L1 listeners, but there is evidence of some individuals with fairly good scores in the auditory condition but random performance on the basis of visual cues. A few Japanese learners with auditory d' around 0 show visual d' of around 1, but this may well be due to random error.

The data were then examined to look at correlations between performance in the auditory and visual modalities. Correlations between performance in the *A* and *V* conditions were also significant for the Korean-L1 ($r=0.768$, $N=42$, $p<0.0001$) and Japanese-L1 ($r=0.477$, $N=78$, $p<0.0001$) groups, suggesting that those learners who showed better consonant identification on the basis of acoustic information also showed better identification on the basis of visual information. As for experiment 1, the data were then examined to look at the degree of *AV* benefit seen for individual learners, by looking at the number of learners for who performance in the *AV* condition was more than a standard deviation away from performance in the A condition. As can be seen in the

right-hand panel of Fig. 3, this was only the case for 5 learners out of 120 (4.2%). This does not exceed the number expected by chance.

F. Discussion

There was no evidence of *AV* benefit for the perception of the /l/-/r/ contrast for either Japanese or Korean learners of English, for the overall group or for the subgroup of lower performers. Korean listeners showed much better identification of the /l/-/r/ contrast than Japanese listeners when presented in audio or audio-visual conditions. Although Korean listeners also showed significantly higher /l/-/r/ identification in the visual condition than Japanese learners, there was a marked discrepancy in Korean learners between their use of auditory and visual information, whereas Japanese learners were generally poor in identifying the contrast using either modality. We had predicted that Korean-L1 learners would have shown better use of visual cues because of their exposure to laterals in their L1, but the poor use of visual information by learners, regardless of their language background and level of proficiency, is probably due to the low degree of visual salience of the contrast. It must be noted though that learners who showed better identification of the contrast in the auditory or audio-visual conditions did show significantly better use of visual cues, so there is some evidence of "attuning" to visual cues with experience.

The effect of L1 background on the perception of the /l/-/r/ contrast was quite marked, with much higher levels of performance achieved by Korean learners. The effect of their greater degree of language instruction (median of 10 years of instruction for Korean learners versus 8 years for Japanese-L1 learners) cannot be discounted, but it must also be noted that this difference in general levels of performance between Korean and Japanese learners of English does mirror previous findings of Ingram and Park (1998), Hardison (2003), and others, and was present when the number of years of English instruction was equated between groups. Ingram and Park suggested that Koreans do have some exposure to alveolar laterals, at least in word-medial position, as these are produced in geminates (liquid in coda position followed by another in onset position), and that this exposure to a [ɾ]-[l] contrast in word-medial position might help in the acquisition of the English /r/-/l/ contrast, at least in word-medial position. In our data, in the auditory condition, this expected difference between performance in word-initial and word-medial position was not obtained. Also, despite the absence of C+liquid clusters in either Korean or Japanese, Koreans showed significantly higher identification rates for consonants in either word-initial or word-medial clusters than in singletons, whereas Japanese learners showed poor performance in both positions and syllable structures. Lee and Lee (2004) also found that the perception of English liquids (audio-only context) by Korean learners was best in word-initial cluster positions.

IV. GENERAL DISCUSSION

This study evaluated the impact of two main factors—visual salience and L1 background—on the use of visual cues in the perception of consonant contrasts by L2 learners of English. The first general finding of this study is the fact that there was either mild or no *AV* benefit in the perception of non-native contrasts for L2 learners across two different contrasts differing in visual distinctiveness, and across learners with different L1s, and thus different relations of L2 contrasts to the phonological system in the L1. This therefore confirms the fact that the provision of visual cues for a second-language learner does not have the same universal "enhancing" effect seen in native speakers, where visual cues can help overcome the effects of hearing impairment, or environmental signal degradation.

Visual salience has an impact on the perception of visual cues to consonant contrasts in both native and non-native languages. Both native speakers and L2 learners of English achieved much poorer scores in the *V* condition for the less salient /l/-/r/ contrast than the highly salient labial/labiodental contrast, for which near-perfect perception was achieved in the *V* condition for native speakers and even some Spanish-L1 learners of English. Our recent phoneme training study with Japanese-L1 listeners on the two contrasts used here (Hazan *et al.*, 2005) showed that audio-visual training was more effective than auditory training for the labial/labiodental contrast but not for the /l/-/r/ contrast. The visual salience of an L2 consonant contrast will therefore be an important factor in the weighting of visual cues for the perception of that contrast.

As predicted, the language background of the learner also affected the use of visual cues in two important ways. First, the relative weighting of visual and auditory cues in the learner's L1 affected the use of visual cues in the L2. Indeed, greater emphasis on auditory than visual cues was evident for Japanese-L1 speakers for both contrasts, and also quite markedly for Korean-L1 speakers, who achieved much higher scores for the /l/-/r/ contrast in the auditory than in the visual condition. A second factor is the specific relation between the native and second language, both at the level of phoneme and viseme inventories. The contrast between labial and labiodental visemes in Spanish seems to have been transferred to English, as shown by near-ceiling effects in the visual-alone condition (experiment 1), even though the specific phonemic contrast between /b/ and /v/ does not occur in Spanish. However, exposure to laterals in Korean did not lead Korean learners of English to make greater use of visual cues to the /l/-/r/ contrast in English.

Significant correlations between scores in the auditory and visual conditions in the two experiments do suggest that increasing proficiency in the use of acoustic-phonetic cues to the non-native contrast is linked with increasing proficiency in the use of visual cues to the contrast. This suggests that, at least for a visually salient contrast, difficulties in the perception of the contrast using visual cues at the early stages of the acquisition of the L2 contrast are due to the difficulty in associating appropriate phoneme labels with each visual cue rather than with difficulties in discriminating between these visual gestures. Results of our training studies (Hazan *et al.*, 2005) also suggest that audio-visual training improves perception in a visual-only condition, so the use of visual cues

might be expected to increase either with general increases in proficiency achieved via language exposure or traditional language teaching, or with targeted training.

Despite these general effects, as in so many studies of L2 acquisition, there is strong evidence of individual variability in L2 perception in learners with similar exposure to the second language (e.g., Bradlow *et al.*, 1997; Gordon *et al.*, 2001), and of individual biases in the use of the auditory and visual modalities. This quite striking degree of individual variability is seen in the crossplots of individual results for the labial/labiodental contrast, with a small number of Japanese-L1 learners achieving ceiling performance in both the *A* and *V* conditions, while most Japanese-L1 learners were at near-chance level in both conditions. Similarly, for the /l/-/r/ contrast, a small number of learners achieved the mean score for native listeners in the *V* condition despite their limited language exposure to spoken English. This variation in "perceptual skill" seen in a minority of L2 learners is not typically addressed in models of L2 acquisition. Studies of the relative weighting of acoustic cues to the /r/-/l/ contrast suggest that individuals vary widely in their reliance on different cues (e.g., Gordon *et al.*, 2001; Iverson *et al.*, 2005). Studies of the relative use of different acoustic cues to a phonemic contrast by native listeners also show that individuals may vary to the degree to which they make use of specific cues (Hazan and Rosen, 1991). The most successful learners are likely to be the ones that give greatest weight to the acoustic cues used by native listeners (primarily *F*3 transitions), as these are likely to be the most consistently informative across speakers and across utterances (Gordon *et al.*, 2001).

In existing models of phonetic perception in second language acquisition, such as SLM (Flege, 1995) and PAM (Best, 1995), perceptual similarity of L2 sounds to L1 sounds strongly influences the predicted relative difficulty of different non-native contrasts. It is recognized that listener-related factors such as age of L2 acquisition and ongoing use of L1 (e.g., Flege *et al.*, 1997; 2003) also strongly affect attained performance in the L2. The basis for the perceptual similarity varies across models: for Best's PAM model, the basis is similarity in articulatory gestures while for Flege's SLM model, it must be measured empirically as it cannot be predicted *a priori*. In order to encompass the use of visual as well as auditory information to L2 phoneme acquisition, models will need to consider the contributions of three further factors influencing the process of acquisition. Our results with Spanish-L1 learners in experiment 1 suggest that the presence of the same visual contrast in the L1 and L2 (even if marking different phonemic contrasts in the two languages) can lead to a transfer of this information into the L2. "Visual similarity" between the L1 and L2 is therefore a factor that may affect L2 phoneme acquisition. Two further factors, which are more language general are the relative weighting of acoustic and visual cues in the L1: learners who attach greater weight to visual information in the L1 are more likely to attend to visual cues to phoneme contrasts in the L2. Finally, just as acoustic salience may be a factor in the ease of acquisition of the contrast (although see Best and McRoberts, 2003), the visual salience of the contrast is likely to affect the degree to which learners attend to visual cues, as seen by differences in performance for the labial/labiodental and /l/-/r/ contrasts.

Finally, these results also suggest that targeted visual training, possibly in a visual-alone condition, could help enhance the use of visual cues by non-native listeners. This could lead to greater perceptual ease in many communicative situations.

ACKNOWLEDGMENTS

Funded by Grant GR/N11148 from the Engineering and Physical Sciences Research Council of Great Britain. We thank Professor M. Taniguchi (Kochi University), Jo Thorp (London Bell School), and the UCL Language Centre for their substantial help in organizing the testing of Japanese students. We also thank Professor Ron Cole and his colleagues for their help in developing the CSLU toolkit-based testing software.

APPENDIX A: STIMULI USED IN EXPERIMENT 1

/pi/, /bi/, /vi/, /pɑ/, /bɑ/, /vɑ/, /pu/, /bu/, /vu/
/ipi/, /ibi/, /ivi/, /ɑpɑ/, /ɑbɑ/, /ɑvɑ/, /upu/, /ubu/, /uvu/
/ip/, /ib/, /iv/, /ɑp/, /ɑb/, /ɑv/, /up/, /ub/, /uv/

APPENDIX B: STIMULI USED IN EXPERIMENT 2

/li/, /ri/, /lɑ/, /rɑ/, /lu/, /ru/
/kli/, /kri/, /klɑ/, /krɑ/, /klu/, kru/
/fli/, /fri/, /flɑ/, /frɑ/, /flu/, fru/
/ili/, /iri/, /ɑlɑ/, /ɑrɑ/, /ulu/, /uru/
/ikli/, /ikri/, /ɑklɑ/, /ɑkrɑ/, /uklu/, /ukru/
/ifli/, /ifri/, /ɑflɑ/, /ɑfrɑ/, /uflu/, /ufru/

Abramson, A., and Lisker, L. (**1972**). "Voice-timing perception in Spanish word-initial stops," Haskins Laboratories Status Report on Speech Research **29/30**, 15–25.

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., and Yamada, T. (**2004**). "Perceived phonetic distance and L2 learning: The case of Japanese /r/ and English /l/ and /r/," J. Phonetics **32**, 233–250

Best, C. (**1995**). "A direct realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, edited by W. Strange (York, Baltimore), pp. 171–204.

Best, C. T., and McRoberts, G. W. (**2003**). "Infant perception of nonnative consonant contrasts that adults assimilate in different ways," Lang Speech **46**, 183–216.

Binnie, C. A., Montgomery, A. A., and Jackson, P. L. (**1974**). "Auditory and visual contribution to the perception of consonants," J. Speech Hear. Res. **17**, 619–630.

Bradlow, A. R., Pisoni, D. B., Yamada, R. A., and Tohkura, Y. (**1997**). "Training Japanese listeners to identify English /r/ and /l/. IV. Some effects of perceptual learning on speech production," J. Acoust. Soc. Am. **101**, 2299–2310.

Ching, Y. C. T. (**1986**). "Voice pitch information for the deaf," in "Towards better communication, cooperation and coordination," Proc. of the First Asia Pacific Regional Conference Deafness, Hong Kong, edited by U. C. L. Engzell, 340–343.

Cole, R. (**1999**). "Tools for research and education in speech science," Proc. Int. Conf. Phon. Sc., San Francisco, 1277–1280.

Flege, J. E. (**1995**). "Second-language speech learning: theory, findings, and problems," in: *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, edited by W. Strange (York, Baltimore), pp. 229–273.

Flege, J. E., Frieda, E. M., and Nozawa, T. (**1997**). "Amount of native language (L1) use affects the pronunciation of an L2," J. Phonetics **25**, 169–186.

Flege, J., Schirru, C., and MacKay, I. (**2003**). "Interaction between the native and second language phonetic subsystems," Speech Commun. **40**, 467–491.
Fowler, C. (**1986**). "An event approach to the study of speech perception from a direct-realist perspective," J. Phonetics **14**, 3–28.
Gordon, P. C., Keyes, L., and Yung, Y. F. (**2001**). "Ability in perceiving nonnative contrasts: Performance on natural and synthetic speech stimuli," Percept. Psychophys. **63**, 746–758.
Grant, K. W., and Seitz, P. F. (**1998**). "Measures of auditory-visual integration in nonsense syllables and sentences," J. Acoust. Soc. Am. **104**, 2438–2450.
Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics.* (Wiley, New York).
Hardison, D. (**1999**). "Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect," Lang. Learn. **49**, 213–283.
Hardison, D. (**2003**). "Acquisition of second-language speech: Effects of visual cues, context, and talker variability," J. Appl. Psychol. **24**, 495–522.
Hardison, D. M. (**2005**). "Variability in bimodal spoken language processing by native and nonnative speakers of English: A closer look at effects of speech style," Speech Commun. **46**, 73–93.
Hazan, V., and Barrett, S. (**2000**). "The development of phonemic categorisation in children aged 6 to 12," J. Phonetics **28**, 377–396.
Hazan, V., and Rosen, S. (**1991**). "Individual variability in the perception of cues to place contrasts in initial stops," Percept. Psychophys. **49**, 187–200.
Hazan, V., Sennema, A., Iba, M., and Faulkner, A. (**2005**). "Effect of audiovisual perceptual training on the perception and production of consonants in Japanese learners of English," Speech Commun. **47**, 360–378.
Hualde, J. I. (**2005**). *The Sounds of Spanish* (Cambridge University Press, Cambridge).
Ingram, J. C. L., and Park, S.-G. (**1998**). "Language, context, and speaker effects in the identification and discrimination of English /r/ and /l/ by Japanese and Korean listeners," J. Acoust. Soc. Am. **103**, 1161–1174.
Iverson, P., Hazan, V., and Bannister, K. (**2005**). "Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults," J. Acoust. Soc. Am. **118**, 3267–3278.
Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., and Siebert, C. (**2003**). "A perceptual interference account of acquisition difficulties for nonnative phonemes," Cognition **87**, B47–B57.
Kuhl, P. K. (**1993**). "Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception," J. Phonetics **21**, 125–139.
Kuhl, P. K., and Meltzoff, A. N. (**1982**). "The bimodal perception of speech in infancy," Science **218**, 1138–1141.
Lee, B., and Lee, S.-H. (**2004**). "Korean Learners' Perception and Production of English Liquids," *Malsori* (Phonetics) the Journal of Phonetic Society of Korea **52**, 61–84.
Lotto, A. J., Sato, M., and Diehl, R. L. (**2004**). "Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/," *From Sound to Sense: 50+ years of discoveries in speech communication*, 11–13 June 2004, MIT, Conference proceedings, edited by J. Slifka, S. Manuel, M. Matthies, C181–C186.
Massaro, D. W. (**1998**). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle* (MIT Press, Cambridge, MA).
Massaro, D. W., Thompson, L. A., Barron, B., and Laren, E. (**1986**). "Developmental changes in visual and auditory contribution to speech perception," J. Exp. Child Psychol. **41**, 93–113.
Massaro, D. W., Tsuzaki, M., Cohen, M. M., Gesi, A., and Heredia, R. (**1993**). "Bimodal speech perception: An examination across languages," J. Phonetics **21**, 445–478.
Mayo, C., and Turk, A. (**2004**). "Adult-child differences in acoustic cue weighting are influenced by segmental context: Children are not always perceptually biased toward transitions," J. Acoust. Soc. Am. **115**, 3184–3194.
McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature (London) **264**, 746–748.
Navarro-Tomás, T. (**1977**). *Manual de pronunciación española* (Handbook of Spanish Pronunciation), 19th ed. Consejo Superior de Investigaciones Científicas (Publicaciones de la Revista de Filología Española, Madrid, Spain).
Nielsen, K. (**2004**). "Segmental differences in the visual contribution to speech intelligibility," Proc. INTERSPEECH-2004, 2533–2536.
Nittrouer, S., and Miller, M. E. (**1997**). "Predicting developmental shifts in perceptual weighting schemes," J. Acoust. Soc. Am. **101**, 2253–2266.
O'Connor, J. D. (**1967**). *Better English Pronunciation* (Cambridge University Press, Cambridge).
Ortega-Llebaria, M., Faulkner, A., and Hazan, V. (**2001**). "Auditory-visual L2 speech perception: Effects of visual cues and acoustic-phonetic context for Spanish learners of English," Proc. AVSP-2001, 149–154.
Patterson, M. L., and Werker, J. F. (**1999**). "Matching phonetic information in lips and voices is robust in 4.5-month-old infants," Infant Behav. Dev. **22**, 237–247.
Patterson, M. L., and Werker, J. F. (**2003**). "Two-month-old infants match phonetic information in lips and voice," Dev. Science **6**, 191–196.
Prieto, P. (**2004**). *Fonètica i Fonologia. Els Sons del Català* (Phonetics and Phonology: the sounds of Catalan). Editorial UOC.
Quilis, A. (**1993**). *Tratato de Fonología y Fonética Españolas* (Manual of Spanish Phonology and Phonetics). (Gredos, Madrid).
Sekiyama, K. (**1997**). "Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects," Percept. Psychophys. **59**, 73–80.
Sekiyama, K., and Tohkura, Y. (**1993**). "Inter-language differences in the influence of visual cues in speech perception," J. Phonetics **21**, 427–444.
Sekiyama, K., Burnham, D., Tam, H., and Erdener, D. (**2003**). "Auditory-visual speech perception development in Japanese and English speakers," Proc. AVSP 2003, 43–47.
Sekiyama, K., Tohkura, Y., and Umeda, M. (**1996**). "A few factors which affect the degree of incorporating lip-read information into speech perception," Proc. ICSLP1996, 1481–1484.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.
Summerfield, Q. (**1983**). "Audio-visual speech perception, lipreading and artificial stimulation," in *Hearing Science and Hearing Disorders*, edited by M. E. Lutman and M. P. Haggard (Academic, London), pp. 131–182.
Walden, B., Prosek, R., Montgomery, A., Scherr, C. K., and Jones, C. J. (**1977**). "Effects of training on the visual recognition of consonants," J. Speech Hear. Res. **20**, 130–145.
Wei, J., Faulkner, A., and Fourcin, A. J. (**1991**). "An application of speech processing and encoding scheme for Chinese lexical tone and consonant perception by hearing impaired listeners," Proc. Eurospeech"91, Genova, Italy.
Werker, J., and Tees, R. (**1984**). "Cross-language speech perception—evidence for perceptual reorganization during the first year of life," Infant Behav. Dev. **7**, 49–63.
Yamada, R. A. (**1995**). "Age of acquisition of second language speech sounds: Perception of American English," in *Speech Perception and Linguistic Experience*, edited by W. Strange (York, Baltimore), pp. 305–320.

The effects of hearing loss on the contribution of high- and low-frequency speech information to speech understanding. II. Sloping hearing loss[a)]

Benjamin W. Y. Hornsby[a)] and Todd A. Ricketts
Department of Hearing and Speech Science, Dan Maddox Hearing Aid Research Laboratory, Vanderbilt University, Vanderbilt Bill Wilkerson Center, Nashville, Tennessee 37232-8242

(Received 17 August 2004; revised 2 December 2005; accepted 2 December 2005)

The speech understanding of persons with sloping high-frequency (HF) hearing impairment (HI) was compared to normal hearing (NH) controls and previous research on persons with "flat" losses [Hornsby and Ricketts (2003). J. Acoust. Soc. Am. **113**, 1706–1717] to examine how hearing loss configuration affects the contribution of speech information in various frequency regions. Speech understanding was assessed at multiple low- and high-pass filter cutoff frequencies. Crossover frequencies, defined as the cutoff frequencies at which low- and high-pass filtering yielded equivalent performance, were significantly lower for the sloping HI, compared to NH, group suggesting that HF HI limits the utility of HF speech information. Speech intelligibility index calculations suggest this limited utility was *not* due simply to reduced audibility but also to the negative effects of high presentation levels and a *poorer*-than-normal use of speech information in the frequency region with the greatest hearing loss (the HF regions). This deficit was comparable, however, to that seen in low-frequency regions of persons with similar HF thresholds and "flat" hearing losses suggesting that sensorineural HI results in a "uniform," rather than frequency-specific, deficit in speech understanding, at least for persons with HF thresholds up to 60–80 dB HL. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2161432]

PACS number(s): 43.71.Ky, 43.66.Ts, 43.66.Sr, 43.71.An [KWG] Pages: 1752–1763

I. INTRODUCTION

Although the negative effect of sensorineural hearing loss (SNHL) on speech understanding is well documented, the benefits and limitations of restoring audibility via amplification continue to be debated. An area that has received recent attention is how the presence and degree of hearing loss, as a function of frequency, affects our ability to make use of amplified speech information. For example, recent research has suggested that, for persons with moderate-to-severe flat SNHL, restoring audibility to low-frequency speech information is more beneficial than restoring audibility to high-frequency speech information, even given a comparable hearing loss in the high-frequencies (Ching, Dillon, and Byrne, 1998; Ching *et al.*, 2001; Hogan and Turner, 1998; Turner and Cummings, 1999; Vickers, Moore, and Baer, 2001). In the studies cited here, speech understanding in quiet was assessed as the frequency range of speech information was systematically varied. The study results generally revealed little or no improvement in speech understanding as *high-frequency* (i.e., above 3000 Hz) speech information was made audible to individuals with SNHL greater than 55–80 dB HL in the high-frequencies. The evidence regarding the effects of hearing loss on low-frequency information is less clear, however, these and other studies have suggested that speech understanding will improve, regardless of degree of hearing loss, as lower frequency information is made available (Turner and Brus, 2001). Taken together these studies suggest that hearing loss results in a "frequency-specific" deficit in speech understanding with the high-frequency regions (>3000 Hz) being most affected.

Factors other than simply degree of high-frequency hearing loss may also affect our ability to make use of amplified high-frequency speech information. Turner and Henry (2002) reported that the presence of background noise may also impact the utility of amplified high-frequency speech information. They measured the speech understanding in noise, of subjects with varying degrees of high-frequency hearing loss, as high-frequency speech information was progressively increased. Although the relative improvements in intelligibility were small when speech information above 3000 Hz was added, persons with hearing loss listening in noise, regardless of their degree of high-frequency hearing loss, were able to use amplified high-frequency information with efficiency comparable to that of persons without hearing loss. This is in contrast to results from an earlier study using the same paradigm but completed in quiet (Hogan and Turner, 1998). The authors suggest that the difference in results obtained in quiet and in noise are due to differences in the relative access to "easy" (i.e., voicing and manner cues) and "more difficult" (i.e., place of articulation) speech cues when speech is presented in quiet versus noise backgrounds.

The absence of functioning inner hair cells in a specified frequency region, recently referred to as a "cochlear dead region," may also affect our ability to make use of amplified speech information (Baer, Moore, and Kluk, 2002; Moore *et al.*, 2000; Summers *et al.*, 2003; Vickers, Moore, and Baer, 2001). It is assumed that information in the acoustic signal

[a)]Electronic mail: ben.hornsby@vanderbilt.edu

 0001-4966/2006/119(3)/1752/12/$22.50

cannot be accurately transmitted to higher auditory centers if there is a lack of functioning inner hair cells (a dead region). Thus restoring audibility to speech information in a dead region is not expected to improve, and may hinder, speech understanding. Baer, Moore, and Kluk, (2002) found that, in noise, persons with hearing loss and dead regions in the high frequencies were less able to make use of amplified high-frequency speech information than persons with hearing loss but without dead regions. Vickers, Moore, and Baer (2001) reported similar findings for persons with and without dead regions tested in quiet. In the two studies cited above the subjects with dead regions tended to have more high-frequency hearing loss than the subjects without dead regions thus differences in audibility between groups may have also played a role in their findings (Rankovic, 2002). Despite this confound, clearly the lack of surviving inner hair cells in a specific area could affect the utility of speech information in that region.

Finally, a recent study by Hornsby and Ricketts (2003) suggested that the configuration of hearing loss may also play a role in determining the utility of amplified speech information. In this study we examined the utility of speech information, as a function of frequency, in individuals with essentially "flat" SNHLs thereby avoiding the confound of variation in degree of hearing loss across frequency. Speech understanding in noise was measured at a variety of low- and high-pass filter cutoff frequencies. These data were used to estimate crossover frequencies for individuals with flat SNHL and a control group of persons without hearing loss. The crossover frequency, defined as the cutoff frequencies at which the low-pass and high-pass filtering yielded equivalent performance, provided an estimate of the contribution of acoustic information in the low- and high-frequency regions to speech understanding (e.g., French and Steinberg, 1947). No significant difference in crossover frequencies was observed between the control group and persons with flat hearing losses.

In addition, speech intelligibility index (SII; ANSI S3.5, 1997) calculations were made to determine if the performance of persons with flat hearing loss could be explained based primarily on audibility of the speech. Although not specifically designed for persons with hearing loss, multiple researchers have used the SII, and its precursor the articulation index, to explore factors other than audibility (e.g., frequency and/or temporal processing abilities, age, cognitive function) that may affect the speech understanding of persons with hearing loss (e.g., Dubno, Dirks, and Schaefer, 1989; Humes, 2002). Results from Hornsby and Ricketts (2003) show that the absolute performance of the flat HI participants was poorer than predicted (see their Fig. 10). However, the relative improvements in performance with changes in filter cutoff frequency were well predicted, suggesting that these listeners utilized amplified speech information, regardless of frequency, in a fashion comparable to persons without hearing loss. The findings of this study suggest that the presence of hearing loss results in a "uniform" rather than frequency-specific deficit in speech understanding. That is the presence of a flat hearing loss would limit the contribution of speech information across all affected frequency regions (resulting in poorer absolute performance but crossover frequencies that are comparable to those of persons with normal hearing) while the presence of a high-frequency hearing loss would limit the contribution of speech information primarily in the high-frequency regions (thus reducing absolute performance for high-frequency speech information and shifting the crossover frequency lower).

The results of Hornsby and Ricketts (2003) are in contrast to some earlier studies focusing on the utility of high-frequency hearing speech information presented in quiet (Ching, Dillon, and Byrne, 1998; Ching *et al.*, 2001; Hogan and Turner, 1998; Turner and Cummings, 1999; Turner and Brus, 2001) and in noise (Amos and Humes, 2001). One potential reason for this difference may be that fundamental differences in the effect of SNHL on speech understanding exist between persons with flat and sloping hearing losses. For example, since persons with high-frequency hearing losses make relatively good use of low-frequency information and speech information is highly redundant across frequencies, persons with high-frequency SNHL may not have to rely heavily on the more degraded high-frequency speech cues. This would be especially true when the speech materials were presented in quiet, as in several previous experiments (e.g., Hogan and Turner, 1998; Ching, Dillon, and Byrne, 1998). In contrast, persons with flat losses having some degradation across all frequencies may require a broader range of cues regardless of the frequency region.

Given the substantial methodological differences between the studies discussed above, it remains unclear whether configuration of hearing loss plays a role in our ability to utilize speech information in various frequency regions, particularly high-frequency information. The primary purpose of this study was to further examine the effect of configuration of hearing loss on the utility of amplified high-frequency speech information in noise. To do this we compare results from subjects with sloping high-frequency hearing losses to participants with normal hearing and to past work using essentially the same methodology in persons with flat losses (Hornsby and Ricketts, 2003).

II. METHODS

A. Participants

A total of 20 participants, 10 with normal hearing (NH) and 10 with hearing loss, participated in this study. All participants with NH passed a pure-tone air conduction screening at 20 dB HL (250–8000 Hz; ANSI S 3.6, 1996) and had no history of otologic pathology. Individuals in the NH group (one male, nine female) ranged in age from 22 to 47 years (mean 28.5).

Ten persons with sloping, high-frequency SNHL (HI group) also participated in this experiment. The average thresholds at 500 and 3000 Hz were 25.5 and 68.0 dB HL, respectively. HI participants ranged in age from 28 to 82 years old (mean 62.9 years). Specific demographic details of the HI participants are provided in Table I.

Auditory thresholds were assessed at octave frequencies between 250–8000 Hz, as well as at the interoctave frequencies of 1500, 3000, and 6000 Hz. HI participants exhibited

TABLE I. Demographic characteristics of the participants with sloping hearing loss.[a]

Subject	Sex	Age	Length of HL (years)	HA use (in years)	Binaural aids	Education	Cause of HL
DD	F	68	20	None	N	15	Presbycusis
JS	M	68	50	None	N	12	Unknown
ET	M	82	12	<1 years	Y	16	Noise/Presbycusis
JT	F	28	22	16 years	Y	19	Congenital
BW	M	58	25	1 years	Y	20	Noise/Presbycusis
AC	F	70	13	9 years	Y	18	Presbycusis
LC	M	67	20	15 years	Y	12	Noise/Presbycusis
RC	M	67	40	4 years	Y	16	Unknown
MC	M	65	14	2 years	Y	12	Noise/Presbycusis
MH	M	56	33	14 years	Y	16	Noise/Unknown
Average		**62.9**	**24.9**	**6.13**		**15.6**	

[a]HL, hearing loss; HA, hearing aids.

essentially symmetrical hearing loss (interaural difference of ⩽15 dB); air-bone gaps ⩽10 dB at all frequencies, and other than hearing loss, reported no history or complaints of otologic pathology, surgery, or unilateral tinnitus. All testing, following the initial threshold assessment, was performed monaurally. The ear chosen for testing was based on (1) limiting the loss at 500 and 4000 Hz to no more than 40 and 80 dB HL, respectively and (2) choosing the ear with the steepest drop in thresholds between 500 and 4000 Hz. In contrast, the test ear was counterbalanced for the NH group. Table II lists the auditory thresholds, of the ear tested, for each of the HI participants. Additional information provided in Table II is described later in the text.

B. Procedures

The procedures followed in this study essentially mirror those of our earlier study (Hornsby and Ricketts, 2003). Sentence recognition in noise, at various filter cutoff frequencies, was assessed for both the NH and HI groups. In addition, both groups completed threshold testing in a speech-shaped background noise and in the broadband "threshold equalizing noise" that accompanies the TEN test (Moore *et al.*, 2000). Participants were compensated for their time on a per session basis.

1. Sentence recognition testing

Sentence recognition was assessed using the connected speech test (CST; Cox, Alexander, and Gilmore, 1987; Cox *et al.*, 1988). The CST uses everyday connected speech as the test material and consists of 28 pairs of passages (24 test and four practice pairs). A total of two passage pairs were completed for each condition, and the score for each condition was based on the average result of these two passage pairs (i.e., based on 100 key words). Sentence recognition was assessed at multiple low- and high-pass filter cutoff frequencies (total of 12 filter conditions) in order to obtain performance versus filter cutoff frequency functions. These functions were used in the derivation of crossover frequency for each subject. As in our previous study, crossover frequencies (defined as the filter cutoff frequency at which the score for low- and high-pass filtered speech is the same) were calculated to allow for the comparison of the relative importance of low- and high-frequency information between

TABLE II. Auditory thresholds (in dB HL) and measured crossover frequencies (in Hz) for participants with sloping hearing loss. In addition, bolded thresholds mark potential dead regions as identified by a positive finding with the TEN test using the less conservative criteria described by Moore *et al.*, 2000 (masked threshold at least 10 dB above quiet threshold and at least 10 dB above the level of the masking noise).

			Frequency (in Hertz)								
Subject	Ear	Crossover frequency	250	500	1000	1500	2000	3000	4000	6000	8000
DD	L	1223	15	25	**25**	**45**	**60**	**60**	**65**	60	60
ET	R	1451	30	25	35	**60**	70	65	65	70	65
JS	L	1033	20	15	25	30	**60**	70	75	85	90+
JT	L	1377	25	30	45	50	55	65	60	60	65
RC	R	1364	30	35	**45**	**55**	**60**	70	70	75	80
MC	L	1128	**15**	20	**30**	**30**	**55**	70	60	**55**	70
AC	R	1198	35	40	55	60	70	70	75	90	95
LC	L	1301	**30**	**20**	**40**	**60**	**65**	**70**	**70**	70	75
BW	L	866	20	20	40	45	50	65	80	90	90
MH	L	1131	15	25	55	65	65	75	80	90	90

groups and thus directly test whether hearing loss results in a frequency-specific deficit in the contribution of speech information.

In all low-pass filter conditions the high-pass filter cutoff frequency was fixed at 178 Hz. Likewise, in all high-pass filter conditions the low-pass filter cutoff frequency was fixed at 7069 Hz. Both groups completed the following 10 filter conditions: low-pass 800 Hz, 1200 Hz, 1600 Hz, 2000 Hz, and 3150 Hz; high-pass 1600 Hz, 1200 Hz, and 800 Hz; wideband (178–7069 Hz); and band-pass (800–3150 Hz). Performance was assessed on each subject in an additional two filter conditions to provide a better indication of the performance versus filter cutoff frequency function for each subject. All testing was completed in two test sessions with at least one CST passage completed in each filter condition during a session.

As in our previous study, the masking noise was a steady state noise filtered to match the long term spectral shape of the CST materials. The speech and noise were digitally mixed at a +6 dB SNR in an attempt to limit ceiling effects among the participants with normal hearing while not causing floor effects for the participants with hearing loss. To improve audibility for the participants with hearing loss, spectral shaping of the mixed speech, and noise stimuli was performed. Subjects in the NH group also listened to this same shaped speech (although generally at a lower level). The spectral shaping was applied to approximate desired sensation level targets (DSL v4.1 software, 1996) for conversational speech, assuming linear amplification for a subject with a sloping high-frequency hearing loss that approximated the average loss of the subjects used in this study. Shaping was verified by measuring the 1/3-octave rms levels of the masking noise (which spectrally matched the speech) in a Zwislocki coupler using a Larson-Davis 814 sound level meter (slow averaging, flat weighting). The rms difference between coupler outputs and DSL targets (250–6000 Hz) was 2.0 dB, with a maximum difference of −4.7 dB. Figure 1 shows the 1/3-octave rms levels (200–6300 Hz) of the CST talker (based on the entire corpus of CST test passages, tracks 32–55), after spectral shaping was applied, and the DSL targets (interpolated to 1/3 octave values) used in this study. Quiet thresholds for the HI subjects are also shown. These data show that, for a few HI participants, quiet thresholds partially limited audibility of high-frequency information (>4000 Hz). This is discussed in more detail below.

The filtered speech and noise were low-pass filtered (10 kHz), amplified and routed to an ER3 insert earphone (Etymotic Research). Levels were calibrated in the wideband condition (178–7069 Hz) in a Zwislocki coupler and no corrections for level were applied in the various filter conditions. Output levels were measured in a Zwislocki coupler using a Larson-Davis 814 sound level meter (C-weighting, slow averaging).

As in our previous experiment, output levels for the wideband speech were 95 dB SPL for the NH participants. This level was selected in an attempt to minimize differences in presentation levels between the NH and HI participants without exceeding loudness discomfort levels of the NH participants. All NH participants in this experiment reported that the 95 dB SPL presentation level was loud, but not uncomfortable. Wideband levels for HI participants varied depending on individual loudness comfort levels. Nine participants preferred an overall speech level of 103 dB SPL while one participant preferred a lower speech level of 94 dB SPL.

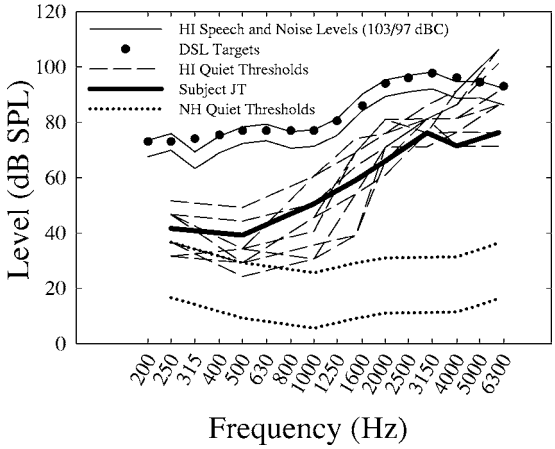


FIG. 1. 1/3 octave band levels (200–6300 Hz) of CST stimuli and DSL targets (filled circles) used in this study. The thin solid lines represent the 1/3 octave band levels of the speech and noise (presented at overall levels of 103 and 97 dBC measured in a Zwislocki coupler). Dotted lines represent a range of thresholds (in dB SPL in a Zwislocki coupler) for our NH control group. These correspond to quiet thresholds of 0 (normal) and 20 dB HL (our screening level for the NH control group). Dashed lines represent quiet thresholds for HI subjects listening to the speech at 103 dBC. The dark solid line represents threshold for HI subject JT, who listened to the speech and noise at a level of 94 and 88 dBC, respectively.

2. Threshold assessment in noise

Monaural masked thresholds, in the same ear used for speech testing, were determined for the octave frequencies of 250–4000 Hz, as well as the interoctave frequencies of 1500, 3000, and 6000 Hz. A clinical method of threshold assessment (ASHA, 1978) was used. However, step sizes were reduced(4 dB down–2 dB up) to improve threshold estimation accuracy. The background noise used during threshold testing was the same noise used during speech testing and was on continuously during threshold assessment. An attempt was made to measure masked thresholds at the same overall masker level as was used during speech testing. Levels, however, varied slightly depending on the loudness tolerance of individual participants. Overall levels of the masking noise, measured in a Zwislocki coupler using a Larson-Davis 814 sound level meter (C-weighting, slow averaging), were 79–89 dB SPL for the NH participants and ranged from 86–97 dB SPL for the HI participants.

3. Diagnosis of dead regions

Cochlear integrity was assessed only for the HI group using the CD version of the "TEN test" (Moore *et al.*, 2000). Pure-tone thresholds for each subject with hearing loss were measured monaurally (same ear as speech testing) in quiet and in the presence of the "threshold equalizing noise," at octave frequencies of 250–4000 Hz and interoctave frequencies of 1500, 3000, and 6000 Hz. TEN levels varied, depending on loudness discomfort issues, from 75–85 dB/ERB

TABLE III. Average SNRs at threshold as a function of frequency for the NH and HI groups, as well as, individual SNRs for participants with sloping hearing loss. Bolded values with asterisks at 6000 Hz represent HI subjects whose quiet thresholds were not shifted by the presence of the masking noise.

	250 Hz	500 Hz	1000 Hz	1500 Hz	2000 Hz	3000 Hz	4000 Hz	6000 Hz	Average
NH	**−2.6**	**−2.8**	**−3.7**	**−9.3**	**−5.4**	**−4.3**	**−3.4**	**0.5**	**−3.9**
					HI participants				
DD	5.2	3.7	5.5	−1.8	2.0	3.7	5.8	5.6	3.4
JS	3.2	3.7	−0.5	−5.8	0.0	−0.3	3.8	21.6	0.6
ET	0.2	0.7	0.5	1.2	1.0	0.7	2.8	**2.6***	1.0
JT	4.2	0.7	4.5	1.2	1.0	−1.3	−3.2	**−1.4***	1.0
BW	10.2	0.7	0.5	−4.8	1.0	0.7	16.8	30.6	3.6
MC	9.2	3.7	5.5	−5.8	0.0	−0.3	3.8	1.6	2.3
RC	1.2	−0.3	−2.5	−5.8	−6.0	−4.3	−2.2	**1.6***	−2.8
LC	13.2	7.7	5.5	2.2	2.0	1.7	5.8	5.6	5.4
AC	5.2	3.7	3.5	−3.8	2.0	−2.3	5.8	**21.6***	2.0
MH	1.2	−0.3	3.5	−1.8	0.0	1.7	7.8	**21.6***	1.7
Average	**5.3**	**2.4**	**2.6**	**−2.5**	**0.3**	**0.0**	**4.7**		**1.8**

(overall levels of 91.5–101.5 dB SPL in a Zwislocki coupler). Moore (2001) suggested using the following criterion to define the presence and extent of a dead region. Specifically, a dead region is identified if (1) the masked threshold is 10 dB or more above absolute threshold in quiet and (2) the masked threshold is at least 10 dB or more above the noise level per ERB.

Recent work by Summers *et al.* (2003), however, questions the specificity of the TEN test, at least when using the Moore's original criterion for identification of dead regions. Summers and colleagues examined the identification of dead regions in persons with high-frequency sloping hearing losses using both the TEN test and psychophysical tuning curves (PTC's). They reported relatively poor agreement between TEN test and PTC results when using the criterion proposed by Moore (2001). When the criterion for identifying a dead region, however, was increased from a masked threshold at least 10 dB above the level of the masking noise to a threshold at least 14 dB above the level of the noise, agreement between TEN test results and PTC results were much improved. Therefore a more conservative criterion (masked thresholds at least 15 dB above the level of the masking noise) was used to identify suspected dead regions among subjects in this study. A level of 15 dB, as opposed to 14 dB, was chosen due to the use of a 5 dB step size, as recommended by Moore (2001), to determine thresholds using the TEN.

III. RESULTS AND ANALYSIS

A. Thresholds in speech noise

As seen in Fig. 1, HI quiet thresholds may have limited high-frequency audibility in some cases. Masked thresholds were compared to thresholds in quiet to confirm that for the most part the masking noise, rather than quiet thresholds, determined audibility differences between groups. Results showed that quiet thresholds for the NH and HI groups were shifted at least 4 dB (five subjects showed shifts $\geqslant$12 dB) by the masking noise at all frequencies tested except 6000 Hz. At 6000 Hz, however, the level of masking noise was not intense enough to cause a 4 dB shift in quiet thresholds for five of the 10 HI participants. For these HI participants, auditory threshold at 6000 Hz rather than the masking noise dictated audibility in noise.

In addition, differences in the "effective masking" of the background noise, between the NH and HI groups, were evaluated by subtracting noise levels (rms level in a 1/3 octave band) from the level of the pure tone at threshold. All measures were made in a Zwislocki coupler. This method provides a comparison of the SNR required for threshold detection in noise (similar to the critical ratio measure) between groups and is provided here to allow for comparison to data reported in the same fashion from our previous study (Hornsby and Ricketts, 2003). Described in this fashion a more negative SNR is better as it suggests that a lower signal SPL is required to detect the signal in noise (e.g., hearing in noise is better). A large positive value suggests that the noise had a significant masking effect or conversely that the masking noise was not intense enough to cause a shift in quiet threshold. The average results for the NH and HI group, as well as individual results for HI participants are shown in Table III for the frequencies of 250–6000 Hz. HI subjects whose quiet thresholds at 6000 Hz were sufficiently poor that they were not shifted by the masking noise are shown bolded and with asterisks.

The data reveal the average SNRs for the HI group were substantially elevated compared to the NH group. When averaged across frequency (250–4000 Hz) the SNR at threshold for the NH participants (−4.5 dB) was 6.3 dB better than that of the HI participants (1.8 dB). Results from 6000 Hz were not included in these averages as quiet thresholds rather than the masking noise were responsible for the SNR for several HI subjects. This difference is comparable to that seen in our previous study (i.e., 5.4 dB difference between groups). In addition, SNRs were elevated compared to the NH group, even in the 250–500 Hz frequency regions where thresholds in quiet for the majority of HI participants were near normal ($\leqslant$25 dB HL) or no poorer than that associated with a mild hearing loss ($\leqslant$40 dB HL).

B. Thresholds in TEN

Although any broadband background noise (e.g., the speech noise described above) may be used to identify suspected dead regions, the TEN test (Moore *et al.*, 2000; Moore, 2001) was specifically designed to allow relatively quick identification and quantification of suspected dead regions. Therefore cochlear integrity was assessed using the TEN. Using a more conservative criteria (i.e., masked threshold at least 10 dB> than quiet and at least 15 dB> than the level of the masker), none of our HI subjects showed a positive result at any test frequency assessed. Using the less conservative criteria proposed by Moore (i.e., masked thresholds only 10 dB above the level of then masking noise), six of the 10 HI subjects in this study had a positive result at a minimum of one test frequency. Threshold frequencies showing a positive result on the TEN test (using the less conservative criteria) are shown in bold in Table II. Of the six HI participants showing positive TEN results, only four showed a positive result at more than one test frequency. In addition several positive results were observed in frequency regions with good hearing. These findings suggest that large scale dead regions were not present in our study subjects.

C. Derivation and examination of crossover frequencies

The method for deriving crossover frequencies is described in detail in our previous paper (Hornsby and Ricketts, 2003). Briefly, two functions were generated for each participant. Each function consisted of average sentence recognition scores (in proportion correct) plotted as a function of the log of their low-pass or high-pass filter cutoff frequency. Then nonlinear regression, using a three-parameter sigmoid function (SPSS, Inc., SigmaPlot V. 5.0), was used to provide a best fit to each data set. These nonlinear regression functions were used to determine the low- and high-pass filter cutoff frequency for which the predicted score was the same (i.e., the crossover frequency). The sigmoid functions for the NH participants provided a good fit to the measured data, with r^2 values ranging from 0.91 to 0.99. The average score at the crossover frequency for the NH group was 50%. The sigmoid function also provided a good fit to the HI data, with r^2 values again ranging from 0.91 to 0.99. The average score (28%) at the crossover frequency, however, was lower for the HI than the NH group.

A single factor between-subjects ANOVA was used to examine differences in crossover frequencies between groups. Significance was defined by an alpha level <0.05 for this, and all other, statistical analyses reported in this paper. The independent and dependent variables were subject group and crossover frequency (in Hz), respectively. Figure 2 shows the mean crossover frequencies for the NH and HI groups as well as individual crossover frequencies for each HI and NH participant.

In contrast to our previous study using subjects with flat losses (Hornsby and Ricketts, 2003), the ANOVA results showed crossover frequencies were significantly ($F_{1,18}=7.96$, $p<0.05$) lower for the HI (1207 Hz) than the NH (1408 Hz) participants. Although the ANOVA results show a significant difference in average crossover frequencies between the NH and HI groups there was considerable overlap between groups. Five of the 10 HI participants had crossover frequencies within the range of the NH subjects. Drawing conclusions about the utility of high-frequency information for persons with hearing loss based on these ANOVA results *alone* would be inappropriate as it is not clear whether the lower crossover frequency of the HI participants is due simply to reduced audibility. These analyses do, however, provide a starting point in examining that question.

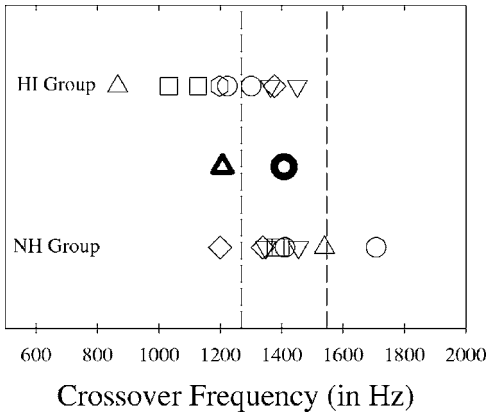


FIG. 2. Crossover frequency data for NH and HI subjects. The dark circle and triangle represent the average crossover frequency for the NH and HI groups, respectively. The remaining symbols represent individual crossover frequencies for the NH (lower) and HI (upper) groups. Dashed lines show plus or minus 1 standard deviation around the NH mean crossover frequency.

Examination of the average speech understanding data for the NH and HI subjects as a function of LP and HP filter condition, as shown in Fig. 3, is useful to help clarify the reason for the lower crossover frequency observed for the HI subjects. As seen in Fig. 3, the speech understanding performance of the HI subjects was substantially poorer in the HP than the LP conditions, compared to the NH control group. Thus the lower-than-normal average crossover frequency seen in the sloping HI group appears to be due primarily to the reduced performance observed in the HP conditions by the HI subjects.

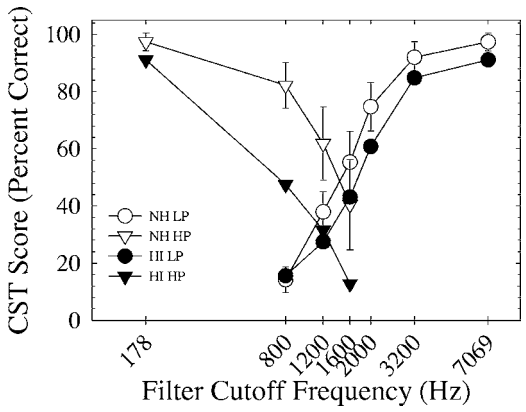


FIG. 3. Average CST scores for NH and HI participants as a function of low- and high-pass filter cutoff frequency. The filled and unfilled symbols show scores for the HI and NH participants, respectively. The circles and triangles represent scores for low- and high-pass filtering, respectively. Error bars show 1 standard deviation around the NH means.

D. Speech intelligibility index (SII) calculations

Results from the crossover frequency data suggest that the NH and HI groups differed in their ability to make use of amplified speech information and that this difference varied with frequency. However, given that quiet thresholds limited high-frequency audibility for some HI subjects and thresholds in noise varied substantially between the NH and HI participants (see Fig. 1 and Table III) it remains unclear whether the observed differences in speech understanding were due to residual differences in audibility or other factors (such as the presence/configuration of hearing loss) resulting in differences in the ability to utilize amplified speech information. Given our subjects had high-frequency hearing loss, reduced audibility of high-frequency speech sounds could provide a simple explanation for the lower-than-normal crossover frequencies observed in the HI group. To examine this question, the 1/3 octave band SII calculations (ANSI S3.5, 1997) were performed to generate SII values and predicted scores in each filter condition for each subject. Measures of quiet thresholds were interpolated or extrapolated to match the 1/3 octave center frequencies used in the SII calculations. These thresholds, along with levels of the speech and noise stimuli, were used to determine the proportion of audible speech information in each 1/3 octave frequency band. The frequency importance function specifically derived for the CST materials was used (Sherbecoe and Studebaker, 2002). In addition, the effective speech peaks for the CST proposed by Sherbecoe and Studebaker (2002) were used, rather than assuming 15 dB peaks as in the ANSI standard. The corrections for high presentation levels and spread of masking effects proposed in the ANSI standard were also included in the calculations. Given our interest in potential differences in performance between the HI and NH groups, however, no corrections for hearing loss desensitization were utilized.

1. SII results for the NH group

Average CST scores for the NH group, as a function of SII are shown in Fig. 4. Each symbol represents the average score of the NH subjects in one of the low-, high-, or band-pass filter conditions previously described. The solid line in Fig. 4 shows the transfer function relating the SII to CST performance. The transfer function was derived, as in our earlier study, using nonlinear regression and a three-parameter sigmoid function (SPSS, Inc.: SigmaPlot V. 5.0) to provide a best fit to the data. For comparative purposes we have also included the transfer function derived by Sherbecoe and Studebaker (2002) for the CST materials (shown by the dotted line). In general, our transfer function required a higher SII for a fixed level of performance (except for high SII values). Given procedural differences in the two studies and the fact that our own transfer function provided a better fit to our data we chose to use our own function.

2. SII predictions of HI crossover frequency data

Utilizing the transfer function derived from the NH data, predictions of individual HI performance based on each person's SII were obtained. These "SII predicted scores" were then compared to actual performance across the various filter conditions. Figure 5 shows the average and individual HI scores as a function of SII and filter cutoff frequency for the low-pass, high-pass, band-pass and wideband conditions.

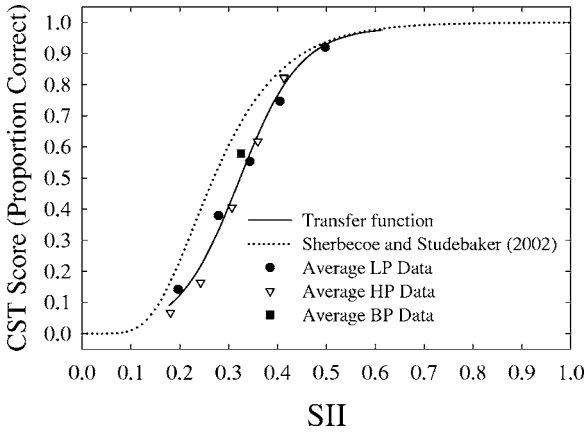


FIG. 4. Transfer function relating SII values to CST performance. This function was derived using data from our NH participants. Average CST scores as a function of SII for our NH group for low-, high-, and band-pass conditions are also shown. For comparative purposes we have also included the transfer function derived by Sherbecoe and Studebaker (2002) for the CST materials (shown by the dotted line).

In addition, SII predicted scores for each HI person were used to derive an "SII predicted crossover frequency" for each subject (as described in Sec. III C above). Figure 6 shows the average predicted and measured results for the HI subjects in several LP and HP conditions and the best fit regression functions to the average data.

The data in Figs. 5 and 6 show that, in general, the SII over-predicted the performance of our HI participants except for conditions where speech information was limited to frequencies below ~1200 Hz (where hearing loss, if present, was less severe). This discrepancy was largest for data obtained in the HP conditions resulting in a lower-than-predicted crossover frequency (Fig. 6). This was confirmed

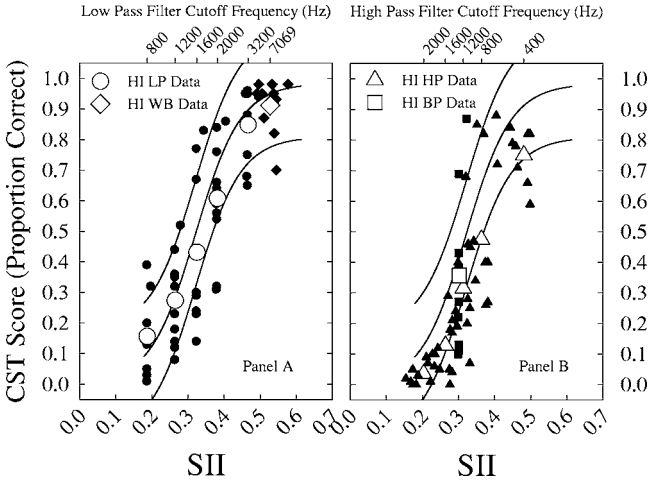


FIG. 5. Average and individual HI CST performance as a function of SII and filter condition. Performance in the low-pass (circles) and wideband (diamonds) conditions are shown in panel A. Band-pass (800–3150 Hz shown by squares) and high-pass (triangles) performance are shown in panel B. Average and individual performance are shown by the open symbols and filled symbols, respectively.

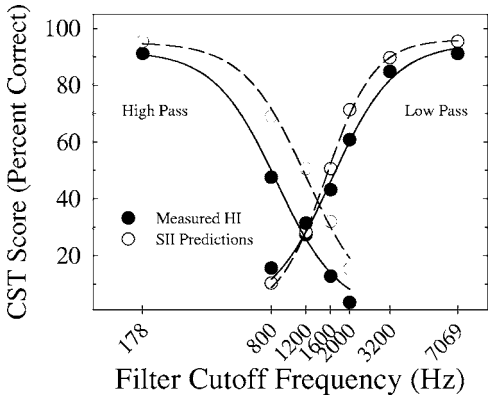


FIG. 6. Average SII predicted and measured CST performance for the HI group in several low- and high-pass filter conditions and the best fit regression functions to the average data. SII predicted and measured performance is shown by the open and filled circles, respectively.

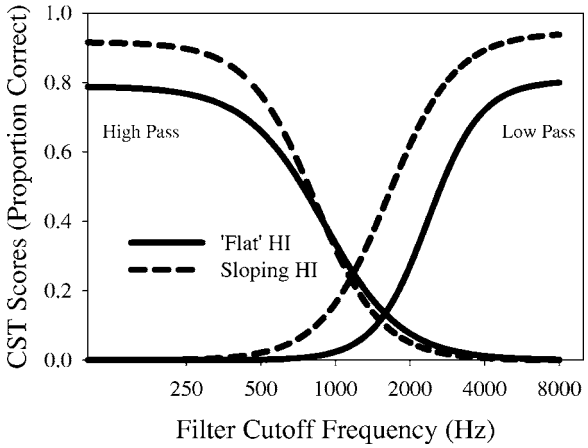


FIG. 7. Average regression functions based on actual measured performance in low- and high-pass filter conditions (i.e., not based on SII predicted scores) of the sloping and flat HI groups. Regression functions for the sloping and flat groups are shown by the dashed and solid lines, respectively.

by an analysis of the individual crossover frequencies derived from the SII predicted scores. Results of a paired-sample *t*-test found the average crossover frequency of 1334.5 Hz, based on SII predicted scores for the HI participants, was significantly higher ($t=-2.897$, $p<0.05$) than the average crossover based on actual HI scores (1207 Hz) but not significantly different ($t=1.699$, $p=0.124$) from the average measured crossover frequency of the participants without hearing loss (1407 Hz).

This finding suggests that the lower crossover frequencies observed in the measured data from our HI participants was not due solely to reduced audibility. Consistent with the earlier comparison to the NH measured data, the lower crossover frequencies of our HI participants appears to be due, at least in part, to poorer-than-normal use of amplified HF information. A primary question of interest however, is whether this poorer performance is substantially larger than that resulting from a comparable low-frequency hearing loss.

To explore this question we compared the average regression function for our sloping HI subjects to that derived from our previous work (Hornsby and Ricketts, 2003) using subjects with a flat hearing loss and essentially equivalent high-frequency thresholds. Figure 7 shows the average regression functions based on actual measured performance in low- and high-pass filter conditions (i.e., not based on SII predicted scores) of the sloping and flat HI groups. Except in the widest bandwidth conditions the regression functions for the high-pass data essentially overlap between flat and sloping groups. In contrast, the regression function for the low-pass data shows consistently poorer performance for the flat HI group resulting in a *lower* crossover frequency for the sloping HI group compared to the flat group. The difference in crossover frequencies, however, is not due to differences in the ability to use high-frequency information. Rather the *higher* crossover frequency observed for the flat hearing loss group is the result of poorer use of low-frequency information. This finding is consistent with the suggestion that hearing loss results in a "uniform" rather than frequency-specific deficit in speech understanding.

3. Effect of presentation level on crossover frequencies

One factor that is well known to negatively affect speech recognition performance is high presentation levels (e.g., French and Steinberg, 1947; Studebaker *et al.*, 1999; Dubno, Horwitz, and Ahlstrom, 2005a, 2005b; Hornsby, Trine, and Ohde, 2005). Given that the speech and noise materials were presented with a high-frequency emphasis in this study (i.e., at a higher-than-normal level) it is possible that the lower-than-predicted crossover frequency observed in our HI participants was due, at least in part, to the higher-than-normal presentation levels of the high-frequency speech. Although we attempted to minimize differences in presentation levels to our NH and HI participants, to limit loudness discomfort our NH participants generally listened at a lower level than our HI participants (95 versus 103 dB SPL).

The SII includes a correction for the negative effects of high presentation levels and thus allows us to estimate the effect of high presentation levels on crossover frequency. To do so we compared predicted crossover frequencies for our NH participants assuming presentation levels of 65, 95, and 103 dB SPL. The regression functions, derived from predicted scores, for each presentation level are shown below in Fig. 8.

The data in Fig. 8 show that predicted crossover frequencies were shifted systematically lower with each increase in presentation level. The lowering was due to the more rapid decline in performance in the high-pass compared to the low-pass conditions. The largest shift occurred with an increase in level from 65 dB to 95 dB SPL with the predicted crossover frequency lowering from 1578 Hz to 1407 Hz. Increasing the level to 103 dB resulted in only a slight additional lowering of the crossover frequency to 1396 Hz. Consistent with past research (e.g., Hogan and Turner, 1998) these findings suggest that spectral shaping appropriate for persons with high-frequency hearing loss results in high presentation levels that limit the utility of amplified high-frequency speech information. To confirm the effect was due to the high level distortion factor within the ANSI standard and not other factors such as spread of mask-

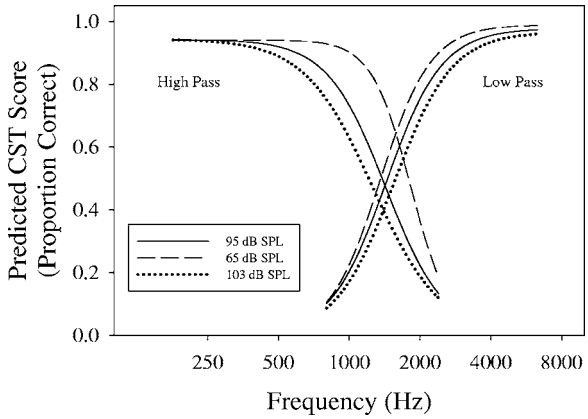


FIG. 8. Effect of presentation level on crossover frequency. The regression functions shown were derived from fits to SII predicted scores for the CST in multiple low- and high-pass filter conditions. Predictions assume the speech and noise were shaped as in the current study for sloping high-frequency hearing loss and presented at the same SNR as used in this study(+6 dB SNR). The functions were derived assuming speech presentation levels of 65, 95, and 103 dB SPL (measured in a Zwislocki coupler), shown by the dotted, dashed, and solid lines, respectively.

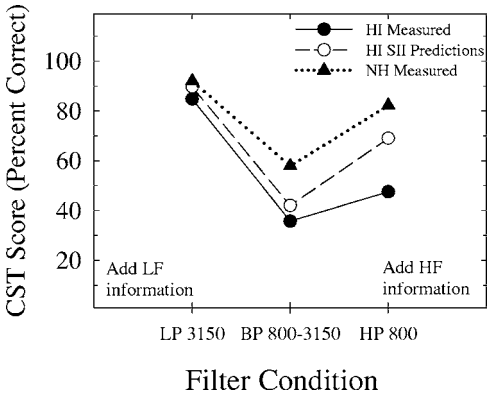


FIG. 9. Average measured and predicted CST scores for the HI group, shown by the filled and open symbols respectively, from the band widening conditions. Measured performance for the NH group (filled triangles) is also shown for comparison.

ing effects which are also included in the calculation, we also derived predicted crossover frequencies without utilizing the level distortion factor in the SII calculations. Without the inclusion of the level distortion factor, predicted crossover frequencies shifted only minimally lower (from 1578 Hz to 1559 Hz) as levels increased from 65 to 103 dB SPL. Likewise, when shaping appropriate for a "flat" hearing loss, as in our previous study, was applied crossover frequencies were only minimally affected (i.e., lowered from 1593 Hz to 1536 Hz).

4. Analysis of band-pass data

The analysis of crossover frequencies provides information about the relative utility of high- and low-frequency information. It is also of interest however, to examine the effect of hearing loss on the utility of speech information in specific frequency regions. We can also use the data in Figs. 3 and 6 to examine the effect of hearing loss on the relative utility of high- and low-frequency speech information. Of particular importance in this study, performance improved for both the NH and HI subjects as high-frequency information above 3000 Hz was made available (i.e., as low-pass filter cutoff frequency increased from 3150 Hz to 7069 Hz). Figure 6 shows that although absolute performance was slightly poorer than predicted, the improvement in performance with the addition of high-frequency information was well predicted by the SII.

It should be noted here that the relative improvement, both measured and predicted, in performance when high-frequency information was added to existing low-frequency information was small (approximately 5%–6% for both groups) for both the NH and HI participants. However, ceiling effects clearly could be limiting performance improvements in this condition. By comparing performance in multiple BP conditions (i.e., BP800-3150, LP3150, and HP800) the contribution of low- and high-frequency information can be assessed while limiting the impact of ceiling effects. These results are shown in Fig. 9.

Absolute performance is only slightly over predicted in the low-pass 3150 and BP 800–3150 Hz conditions. In addition, the performance improvement as additional low-frequency information is made available is well predicted by the SII. In contrast, when additional high-frequency information is provided (compare BP800–3150 Hz to HP800 Hz) the performance improvement is somewhat less than predicted(~12% versus 27%). However, it is important to note that this limited improvement with the addition of high-frequency information coincides with the frequency regions where the hearing loss is most severe. In other words, this reduced use of high frequency information is expected given the presence of hearing loss in this frequency region. It is not evidence, however, that the high-frequency regions are affected any more than low frequency regions given a comparable degree of hearing loss. In fact when viewed in conjunction with the data in Fig. 7 (performance of flat and sloping HI groups), which shows comparable performance in the high pass conditions, this finding is consistent with the idea that the presence of hearing loss limits the utility of speech information in a uniform fashion across the frequency region of the hearing loss.

IV. DISCUSSION

One purpose of this study was to examine the relative utility of amplified high-frequency speech information for persons with sloping high-frequency hearing loss. To do this we examined the speech understanding of persons with sloping high-frequency hearing loss under various conditions of filtering and compared their performance to that of a control group with normal hearing and to previous work using a person with flat hearing losses. It is important to note that, on average, persons with sloping hearing loss were able to use amplified high-frequency information to improve their speech understanding, although the magnitude of the improvement was small. The sloping HI subjects' performance improved an average of 6% (compared to the NH improvement of 5%) as LP filter cutoff frequency increased from

3150 Hz to 7069 Hz with eight of 10 HI subjects showing improvements of 7% or less. This was due to in large part to ceiling effects as average performance in the 3150 Hz low-pass condition was ~92%. Two of the 10 sloping HI subjects, AC and ET, showed larger improvements of 10% and 27%, respectively. These data show that some individuals with sloping high-frequency hearing loss obtain substantial benefit from speech information above 3000 Hz. These results are consistent with the findings of Turner and Henry (2002), whose HI participants were also tested in noise. Together these findings suggest that although persons with sloping high-frequency hearing loss (up to 60–80 dB HL) are able to make use of amplified high-frequency speech information, the improvements in most cases are expected to be small.

The average improvement in scores as LP filter cutoff increased above 3150 Hz was larger for the subjects with flat hearing loss in our previous study than for those with sloping losses in the current study, even though high-frequency thresholds were comparable between groups. Subjects with flat losses showed an average improvement of ~16% as low-pass filter cutoff frequency increased from 3150 Hz to the wideband condition (Hornsby and Ricketts, 2003). The range of improvement was again variable, although five of nine subjects showed improvements of 17% or more (maximum of 33% improvement). This larger improvement was likely due, at least in part, to reduced ceiling effects. Average performance in the low-pass 3150 Hz condition was only ~62% for the flat hearing loss group. In addition, and in contrast to some reports in the literature (e.g., Hogan and Turner, 1998; Vickers, Moore, and Baer, 2001), none of our subjects in either study showed any decrement in performance as additional high-frequency information was added.

These cases highlight the utility of amplified high-frequency information when provided in conjunction to existing low-frequency information and are most applicable in relation to hearing aid fittings. These data argue for a cautious approach when considering reducing high-frequency audibility (at least for speech in noise) as the majority of our subjects in both studies showed improvements in speech understanding in noise when the maximum bandwidth of speech information was made available. At the same time, both clinical experience and research suggests that in addition to speech intelligibility, we must take into account an individual's perception of loudness comfort and sound quality, which may be negatively affected by substantial high-frequency amplification, when determining optimal amplification characteristics, particularly for adults.

A primary purpose of this study was to examine the role hearing loss configuration plays in the relative utility of amplified high-frequency speech information. Several findings from this study suggest that persons with sloping high-frequency SNHL do not utilize amplified speech information in the same fashion as persons with normal hearing. For example, crossover frequencies, derived from measured speech scores in low- and high-pass filter conditions, were significantly lower for the persons with sloping SNHL than for the control group of persons without hearing loss. The average measured crossover frequency, for the sloping HI group, was also lower than the average SII predicted crossover frequency for these same subjects. This suggests that these lower-than-normal crossover frequencies were not due simply to reduced audibility. In addition, when ceiling effects were limited, as in the band widening conditions, our sloping HI group showed smaller improvements in speech understanding with the addition of high-frequency information than the NH participants (12% vs 24% improvement) and smaller improvements than predicted based on SII calculations (12% vs 27%).

It is not clear however, what factors may be responsible for this limited utility of amplified high-frequency information. Specifically, factors other than degree of hearing loss also differentiate our NH and sloping hearing loss groups and may have impacted our results. In addition to hearing loss, a substantial difference in age exists between our NH and HI groups (~29 and 63 years old, respectively). Other researchers have reported factors, besides hearing loss, associated with the aging process (e.g., age related declines in processing efficiency, memory and cognitive processing) can influence speech understanding in elderly (CHABA, 1988; Wingfield, 1996; Amos and Humes, 2001; Pichora-Fuller, 2003). Amos and Humes (2001) reported specifically that some measures of cognitive function were correlated with the ability of elderly HI persons with varying degrees of high-frequency hearing loss, to make use of high-frequency speech information, although this contribution was secondary to audibility. The fact that our HI participants showed substantially higher thresholds in noise compared to the NH participants, even in the lower frequency regions (250–500 Hz) where thresholds were best, strongly suggests that factors other than hearing loss affected our results.

At the same time a relationship between speech understanding and age-related processing deficits is not always present or clear (e.g., Humes and Floyd, 2005). In the current study comparing variations in the accuracy of SII predictions across frequency regions with and without hearing loss provides some estimate of the impact of age on our results. Note in Fig. 6 that performance in the low-pass conditions ($<$1200 Hz), where hearing thresholds are better, is well predicted. In contrast, whenever, speech information is presented to frequency regions with more hearing loss (e.g., $>$1200 Hz) performance is generally poorer than predicted. If age differences were a substantial factor in our findings we would expect that predictive accuracy would have been affected across all frequency regions. Finally, the presence or absence of age effects in the current study do not detract from the importance of the primary finding suggesting that hearing loss (up to ~80 dB HL) has a "uniform" rather than "frequency-specific" effect on speech understanding.

In contrast to some previous studies (e.g., Vickers, Moore, and Baer, 2001; Baer, Moore, and Kluk, 2002) our study results suggest that cochlear dead regions were not a significant factor in limiting the utility of high-frequency speech information. In addition, we used SII calculations to assess the role audibility played in the observed differences between the groups. SII calculations suggest that although slight audibility differences between the NH and HI groups

existed in the high frequencies, these differences had only a minimal impact on the lowered crossover frequencies observed for our sloping HI participants.

Likewise, differences in absolute presentation levels between groups (the HI listened at slightly higher levels) appeared to play only a small role in the poorer use of amplified high-frequency information compared to our NH control group. It should be noted, however, that when compared to predicted performance at conversational levels (e.g., 65 dB SPL) the high presentation levels required by persons with high-frequency hearing loss to ensure audibility appear to substantially reduce the utility of high-frequency speech information both for listeners with and without hearing loss. This is evidenced by the lowering in predicted crossover frequencies from 1578 to 1396 Hz as presentation levels are increased from 65 to 103 dB SPL. This is also consistent with the difference in the measured crossover frequencies obtained by Sherbecoe and Studebaker (2002) for the CST materials presented at 65 dB SPL in $a+6$ dB SNR (1562 Hz) and that of the NH participants in the current study (1407 Hz). Even when these factors were accounted for in the SII calculations (i.e., comparing predicted and measured HI crossover frequencies), however, the utility of high-frequency information remained poorer-than-normal for these persons with sloping high-frequency hearing loss.

We also looked at the effect of hearing loss configuration by comparing our results from persons with sloping hearing loss to previous data from persons with flat configuration. The comparison of crossover frequencies (see Fig. 7) between the sloping HI group from the current study and the flat hearing loss group from our previous study (Hornsby and Ricketts, 2003) suggests that configuration of hearing loss does play a role in determining the relative utility of speech information in different frequency regions. The data in Fig. 7, however, suggest that the presence of hearing loss results in a uniform rather than frequency-specific deficit in speech understanding. Specifically, the difference in crossover frequencies between the sloping HI and flat HI groups was due to the poorer performance in the *low-pass* conditions by the flat HI group (i.e., in the region where thresholds were poorer for the flat compared to sloping group). In addition crossover frequencies for our flat group and a control group of persons without hearing loss were not significantly different. This suggests that presence of a flat hearing loss decreased the utility of speech information across frequencies in a "uniform" fashion. The introduction of a sloping configuration merely resulted in improved utility of low-frequency information (where thresholds were better) and continued, and comparable, poorer use of high-frequency information (where thresholds remained poor).

V. CONCLUSIONS

A primary finding of this study is that persons with sloping high-frequency SNHL are limited in their ability to make use of audible high-frequency speech information to improve speech understanding in noise. This deficit appears due in part to the relatively high presentation levels these individuals are forced to listen. Audibility measures (i.e., SII calculations) suggest that the limited utility of high-frequency information by persons with sloping hearing loss in this study is *not* due simply to reduced audibility. Rather, SII measures suggest that the presence of sensorineural hearing loss, regardless of frequency, results in a *poorer*-than-normal ability to use amplified speech information in that frequency region. This deficit appears to be comparable to that seen in persons with flat hearing loss configuration and comparable high-frequency thresholds, suggesting configuration of hearing loss does not create *additional* reductions in the utility of amplified high-frequency information. Taken together these findings add further support to the idea that sensorineural hearing loss (up to 60–80 dB HL) reduces the contribution of amplified speech information in a uniform, rather than frequency-specific, fashion at least for high-frequency (3000–4000 Hz) hearing losses up to approximately 60–80 dB HL.

ACKNOWLEDGMENTS

This work was supported, in part, by Grant R03 DC006576 from the National Institute on Deafness and Other Communication Disorders (National Institutes of Health) and the Dan Maddox Hearing Aid Research Foundation. Thanks to the associate editor Ken Grant, Larry Humes, and two anonymous reviewers for their helpful comments on an earlier version of this paper. Thanks also to Gerald Studebaker and Bob Sherbecoe for their patient discussions regarding the SII procedure and to Earl Johnson for help with data collection. This research was supported, in part, by the Dan Maddox Hearing Aid Research Endowment.

Amos, N., and Humes, L. (**2001**). The contribution of high frequencies to speech recognition in sensorineural hearing loss. Paper presented at the 12th International Symposium on Hearing, Mierlo, The Netherlands.

ANSI. (1996). ANSI S3.6-1996 American National Standard Specification for Audiometers, New York.

ANSI. (1997). ANSI S3.5-1997 American National Standard Methods for the calculation of the speech intelligibility index, New York.

ASHA. (**1978**). "Manual pure-tone threshold audiometry," **20**, 297–301.

Baer, T., Moore, B. C., and Kluk, K. (**2002**). "Effects of low pass filtering on the intelligibility of speech in noise for people with and without dead regions at high frequencies," J. Acoust. Soc. Am. **112**, 1133–1144.

CHABA. (**1988**). "Speech understanding and aging. Working group on speech understanding and aging. Committee on hearing, bioacoustics, and biomechanics, Commission on behavioral and social sciences and education, National Research Council," J. Acoust. Soc. Am. **83**, 859–895.

Ching, T., Dillon, H., and Byrne, D. (**1998**). "Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification," J. Acoust. Soc. Am. **103**, 1128–1140.

Ching, T. Y., Dillon, H., Katsch, R., and Byrne, D. (**2001**). "Maximizing effective audibility in hearing aid fitting," Ear Hear. **22**, 212–224.

Cox, R. M., Alexander, G. C., and Gilmore, C. (**1987**). "Development of the connected speech test (CST)," Ear Hear. **8**, 119S–126S.

Cox, R. M., Alexander, G. C., Gilmore, C., and Pusakulich, K. M. (**1988**). "Use of the Connected Speech Test (CST) with hearing-impaired listeners," Ear Hear. **9**, 198–207.

Dubno, J. R., Dirks, D. D., and Schaefer, A. B. (**1989**). "Stop-consonant recognition for normal-hearing listeners and listeners with high-frequency hearing loss. II-Articulation index predictions," J. Acoust. Soc. Am. **85**, 355–364.

Dubno, J., Horwitz, A., and Ahlstrom, J. (**2005a**). "Word recognition in noise at higher-than-normal levels: Decreases in scores and increases in masking," J. Acoust. Soc. Am. **118**, 914–922.

Dubno, J., Horwitz, A., and Ahlstrom, J. (**2005b**). "Recognition of filtered words in noise at higher-than-normal levels: Decreases in scores with and without increases in masking," J. Acoust. Soc. Am. **118**, 923–933.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.
Hogan, C. A., and Turner, C. W. (**1998**). "High-frequency audibility: benefits for hearing-impaired listeners," J. Acoust. Soc. Am. **104**, 432–441.
Hornsby, B. W., and Ricketts, T. A. (**2003**). "The effects of hearing loss on the contribution of high- and low-frequency speech information to speech understanding," J. Acoust. Soc. Am. **113**, 1706–1717.
Hornsby, B. W., Trine, T., and Ohde, R. (**2005**). "The effects of high presentation levels on consonant feature transmission," J. Acoust. Soc. Am. **118**, 1719–1729.
Humes, L. E. (**2002**). "Factors underlying the speech-recognition performance of elderly hearing-aid wearers," J. Acoust. Soc. Am. **112**, 1112–1132.
Humes, L. E., and Floyd, S. S. (**2005**). "Measures of working memory, sequence learning, and speech recognition in the elderly," J. Speech Lang. Hear. Res. **48**, 224–235.
Moore, B. C., Huss, M., Vickers, D. A., Glasberg, B. R., and Alcantara, J. I. (**2000**). "A test for the diagnosis of dead regions in the cochlea," Br. J. Audiol. **34**, 205–224.
Moore, B. C. J. (**2001**). "Dead regions in the cochlea: diagnosis, perceptual consequences, and implications for the fitting of hearing aids," Trends Amplification **5**, 1–34.
Pichora-Fuller, M. K. (**2003**). "Processing speed and timing in aging adults: psychoacoustics, speech perception, and comprehension," Int. J. Audiol. **42**, S59–S67.
Rankovic, C. M. (**2002**). "Articulation index predictions for hearing-impaired listeners with and without cochlear dead regions," J. Acoust. Soc. Am. **111**, 2545–2548; **111**, 2549–2550 (2002), author reply.
Sherbecoe, R. L., and Studebaker, G. A. (**2002**). "Audibility-index functions for the connected speech test," Ear Hear. **23**, 385–398.
Studebaker, G., Sherbecoe, R., McDaniel, D., and Gwaltney, C. (**1999**). "Monosyllabic word recognition at higher-than-normal speech and noise levels," J. Acoust. Soc. Am. **105**, 2431–2444.
Summers, V., Molis, M. R., Musch, H., Walden, B. E., Surr, R. K., and Cord, M. T. (**2003**). "Identifying dead regions in the cochlea: psychophysical tuning curves and tone detection in threshold-equalizing noise," Ear Hear. **24**, 133–142.
Turner, C. W., and Brus, S. L. (**2001**). "Providing low- and mid-frequency speech information to listeners with sensorineural hearing loss," J. Acoust. Soc. Am. **109**, 2999–3006.
Turner, C. W., and Cummings, K. J. (**1999**). "Speech audibility for listeners with high-frequency hearing loss," Am. J. Audiol. **8**, 47–56.
Turner, C. W., and Henry, B. A. (**2002**). "Benefits of amplification for speech recognition in background noise," J. Acoust. Soc. Am. **112**, 1675–1680.
Wingfield, A. (**1996**). "Cognitive factors in auditory performance: context, speed of processing, and constraints of memory," J. Am. Acad. Audiol **7**, 175–182.
Vickers, D. A., Moore, B. C., and Baer, T. (**2001**). "Effects of low-pass filtering on the intelligibility of speech in quiet for people with and without dead regions at high frequencies," J. Acoust. Soc. Am. **110**, 1164–1175.

Constrained tone transformation technique for separation and combination of Mandarin tone and intonation

Jinfu Ni[a] and Hisashi Kawai[b]
ATR Spoken Language Communication Research Laboratories, 2-2-2 Hikaridai, "Keihanna Science City" Seika-cho, Kyoto 619-0288, Japan

Keikichi Hirose
Department of Information and Communication Engineering, Graduate School of Information Science and Technology, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

(Received 22 March 2005; revised 16 December 2005; accepted 16 December 2005)

This paper addresses a classical but important problem: The coupling of lexical tones and sentence intonation in tonal languages, such as Chinese, focusing particularly on voice fundamental frequency (F_0) contours of speech. It is important because it forms the basis of speech synthesis technology and prosody analysis. We provide a solution to the problem with a constrained tone transformation technique based on structural modeling of the F_0 contours. This consists of transforming target values in pairs from norms to variants. These targets are intended to sparsely specify the prosodic contributions to the F_0 contours, while the alignment of target pairs between norms and variants is based on underlying lexical tone structures. When the norms take the citation forms of lexical tones, the technique makes it possible to separate sentence intonation from observed F_0 contours. When the norms take normative F_0 contours, it is possible to measure intonation variations from the norms to the variants, both having identical lexical tone structures. This paper explains the underlying scientific and linguistic principles and presents an algorithm that was implemented on computers. The method's capability of separating and combining tone and intonation is evaluated through analysis and re-synthesis of several hundred observed F_0 contours. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2165071]

PACS number(s): 43.72.Ar, 43.72.Ja, 43.70.Fq [DOS] Pages: 1764–1782

I. INTRODUCTION

In spoken Chinese, the voice fundamental frequency (F_0) contours of speech simultaneously manifest syllabic tones and sentence intonation. In Mandarin, four lexical tones exist, traditionally called the first, second, third, and fourth tones (henceforth, Tones 1 to 4), each of which has its own features distinct from others as well as a neutral tone (Tone 0) that has no distinctive feature. For example, there are five distinct words with the same phonemic string: *ma1* (mother), *ma2* (numb), *ma3* (horse), *ma4* (to curse), and *ma0* (a question particle); each is distinguished from the others by a unique tone type. Meanwhile, a sentence, for example, can be uttered as a question or a statement; and a part of the utterance can be emphasized (Xu, 1999). Moreover, the attitudes and emotions of the speaker can be added to the cognitive meaning of the words (Chao, 1968). Thus, Mandarin has intonation-conveying communicative functions similar to nontonal languages, such as English, as described in ToBI systems (Silverman *et al.*, 1992). This poses an important problem: How to deal with the acoustic correlates of lexical tones and sentence intonation. It is important because it is the basis of speech synthesis technology (van Santen *et al.*, 1998). In speech communication systems (O'Shaughnessy, 1987), a text-to-speech (TTS) system is a necessary component. When applying TTS technology, as described in Kawai *et al.* (2004), to a conversational system, for example, synthetic speech must have certain communicative functions in the next generation.

There have been a number of studies focusing on this problem, based on approaches with different emphases. Chinese linguists used to analyze pitch ranges of syllables (Shen, 1985) or phrases (Wu, 1996) considering the tone contexts involving both low and high target points. When all of the syllabic tones in an utterance happen to be of Tone 1, the measurement of pitch ranges may suffer shortcomings. This is because Tone 1 has high-level characteristics; therefore, no low target point is available to give a reference to this measurement. Solutions to the problem may be found, for example, in Garding's (1987) treatment of sentence intonation as appropriate combinations of grids to which the movement of syllabic tones is limited, and the work of Kochanski *et al.* (2003), which uses Stem-ML discussed in depth in Kochanski and Shih (2003) to describe the interactions between nearby tones organized as soft templates. Solutions may also be found in Hirose *et al.* (1994), using the Fujisaki model (Fujisaki and Hirose, 1984) to represent an F_0 contour as a sum of tone and phrase components with a base value on the logarithmic scale of fundamental frequency. In addition, there exist other techniques for representing the relationships between the acoustic and linguistic features, such as rules-based methods (Lee *et al.*, 1993), statistics-based

[a]Electronic mail: jinfu.ni@atr.jp
[b]Address for correspondence: KDDI R&D Laboratories Inc., Speech Processing Laboratory, 2-1-15 Ohara, Kamifukuoka, Saitama, 356-8502, Japan.

approaches to modeling the effects of multiple affecting factors on the F_0 contours (Chen *et al.*, 2005), machine learning techniques, like CART tree models (Dusterhoff *et al.*, 1999) within the Tilt intonation model of Taylor (2000), dynamical system models (Ross and Ostendorf, 1999), hidden Markov models based prosody generation (Tokuda *et al.*, 2000), etc.

This paper presents a novel approach to the quantitative separation of lexical tones and sentence intonation from observed F_0 contours. This consists of transforming *target* values in pairs from *norms* (citation values) to *variants* (observed values) using a constrained tone transformation technique based on structural modeling of the F_0 contours. These targets are intended to sparsely specify the prosodic contributions (of syllabic tones and sentence intonation) to the F_0 contours. The alignment of target pairs between the norms and variants is based on the underlying lexical tone structures. This consists of assuming that a lexical tone structure underlying an F_0 contour actually constructs a framework whereby both syllabic tones and sentence intonation acoustically couple together. Citation values of the targets that define the elements of lexical tone structures are specified by those measured from the citation forms of lexical tones. Theoretically, there exist only four citation forms of lexical tones in a voice, and the four citation forms can be determined by averaging several isolated syllable samples produced in the voice. Accordingly, by suppressing the effects of lexical tones on the F_0 contours, the outcome of this approach is extracted intonation components. Because the lexical tone structures specified by the citation forms of lexical tones also have the function of distinguishing word meaning, this technique thus makes it possible for researchers to focus on studying the relationships between the intonation components and high-level linguistic features rather than the low-level linguistic features (lexical tones). Here, we make a number of specific proposals as follows for implementing this approach on a computer: (1) An approach toward deriving a series of targets from an observed F_0 contour; (2) an algorithm for extracting intonation components from these targets; (3) an algorithm for coupling tone and intonation into a series of targets, given the temporal structures; and (4) a program for computing the transitions through these targets to generate F_0 contours.

This paper is organized as follows. Section II introduces the background assumptions, then Sec. III describes the algorithms and explains the scientific and linguistic principles behind them. Section IV presents the experimental results and discussions. Finally, Sec. V concludes the paper.

II. BACKGROUND ASSUMPTIONS

The F_0 contours of utterances basically represent the acoustic correlates of syllabic tones and sentence intonation in Mandarin. Such a representation obviously contains much redundancy in that an F_0 contour is continuous but linguistic information conveyed by the F_0 contour is discrete. In order to eliminate this redundancy as a first step toward a sparser and structural representation, instead of the acoustic representation, we intend to set up some assumptions drawing upon findings in the literature on tone and intonation.

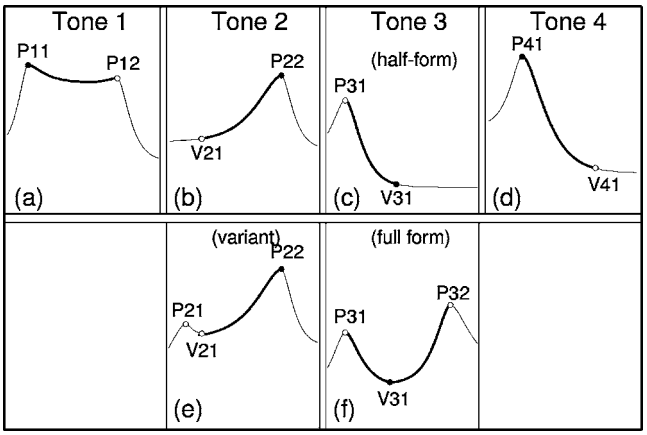


FIG. 1. Illustration of tone graphs sparsely specified by peak and valley targets (circles); the black circles indicate primary targets; and the thick lines indicate the nucleus of a tone.

A. Sparser specification of F_0 contours

It is straightforward to analyze the prosodic contribution (of syllabic tones and sentence intonation) to the F_0 contours as a series of target points, focusing particularly on tonal F_0 peaks and valleys, henceforth referred to as *targets*. Such a target-based sparser specification of F_0 contours is evident in English (Pierrehumbert, 1981; Hirst *et al.*, 2000). In Chinese, this may still be a theoretical issue (if not just an empirical question) without general consensus on whether the F_0 contours can be sparsely represented by a series of targets. We shall assume this kind of representation in the work because the assumption places a strong constraint on the nature of sparser representation of the acoustic correlates of Mandarin tone and intonation. In fact, Garding (1987) expressed Chinese tones with respect to grids (for sentence intonation) as highs and lows or as combinations of highs and lows, which become turning points in the final F_0 contours. In addition, there is no doubt that Mandarin tones are represented by a few pitch targets on a five-point scale: *55* (Tone 1), *35* (Tone 2), *214* (Tone 3), and *51* (Tone 4), where *1* indicates low; *2*, half-low; *3*, middle; *4*, half-high; and *5*, high in a vocal range (Chao, 1968).

A lexical tone is of a graphic structure. Chao (1968) used the term *tone graph* to indicate its abstract characteristics: high-level (Tone 1), high-rising (Tone 2), low-dipping (Tone 3), and high-falling (Tone 4). To specify the elements of this structure by simple targets, we map the five pitch targets *1* to *5* to either peak or valley targets. More specifically, both *4* and *5* are mapped to peaks, and *1* to a valley. For either *2* or *3*, if followed by a lower pitch target, it is regarded as a peak. Otherwise, it is a valley. Figure 1 illustrates a model of tone graphs specified by a series of targets (white and black circles). Pkj (Vkj) indicates the jth peak (valley) target of the kth tone, $k=1, 2, 3, 4$; and $j \in \{1,2\}$. Hereafter, these categorized targets are called *tonal targets*. Tonal targets define the *elements* of structures of lexical tones. Black circles (or other symbols relating to tone graphs throughout this paper) indicate *primary targets* whose F_0 values within prominent syllables are higher for Tones 1, 2, and 4, but lower for Tone 3 (Kratochvil, 1998). The transitions (continuous lines) between targets are approximated by

a second-order filter described in the next section. There is no extra tone graph for Tone 0, because it has no distinctive feature in itself. It may, however, share a part of a tone graph with the preceding tone. Note that these tone graphs have proven to be sufficient for generating almost all the observed tone patterns in spoken Mandarin (Ni *et al.*, 2000, 2003).

The surface tones manifested by F_0 contours may be different from underlying lexical tones due to *tone sandhi*. A well-known tone sandhi rule is that Tone 3 changes to Tone 2 when it is followed by another Tone 3 (Chao, 1968). We use the term *lexical tone structure* to indicate the surface form of a sequence of syllabic tones subject to appropriate tone sandhi rules. A lexical tone structure is thus specified by the tone graphs of individual tones. The elements (i.e., tonal targets) of tone graphs then define the *elements* of lexical tone structures. When focusing on the elements of lexical tone structures, two additional rules are given as follows: (1) Tone 3 takes its half-form (i.e., *21*) when Tone 3 is closely followed by any tone except for Tone 3 (Chao, 1968), and (2) tone 2 takes the variant shown in Fig. 1(e) when it is followed by either Tone 4 or half-Tone 3, each having a low target at the end.

Besides syllabic tones, an F_0 contour also manifests sentence intonation. An intonation theory advanced by Kratochvil (1998) states that, within the prosodic complex of an utterance, the smallest distinctive configuration is tone, while larger configurations up to and including the complex itself are specific modifications of a string of one or more tones. The larger configurations can be predicted given the string and modifications, whereas they cannot be predicted from the modifications alone (Kratochvil, 1998). Thus, a lexical tone structure shall actually provide a construction for coupling both syllabic tones and sentence intonation together with the temporal structures of utterances to yield the speech melody.

As a summary of the discussions above, we set up the following assumptions:

(1) The prosodic contributions to F_0 contours can be analyzed as a series of targets without losing the information on tone and intonation conveyed by the F_0 contours.
(2) The number and types of these targets directly correspond to those of the elements of lexical tone structures underlying the F_0 contours.

An analysis and perceptual experiments are conducted to justify the two assumptions in Sec. IV C.

B. A functional model

A functional model (hereafter, the model) proposed by Ni and Hirose (2000) is employed here to represent the F_0 contours. The model involves two levels of (local and global) control mechanisms intricately joined together. A lexical tone structure underlying an F_0 contour is specified by a sequence of tone graphs hypothetically independent of individual voices and sentence intonation. The heights and ranges of these tone graphs in the frequency-time space are then modulated through compressing or expanding them according to sentence modality. This is done by manipulating the damping factor of a second-order filter derived from a forced-vibrating-based physical interpretation of the body-cover hypothesis of vocal folds (Hirano, 1976), which is seen as a more quantitative look at the mechanics of F_0 control, as described in Titze (1997).

1. Global control

Let $F_0(t)$ represent an F_0 contour as a function of time t, and f_{0_b} and f_{0_t} indicate the bottom and top frequencies of a speaker's vocal range, respectively. Let $\Lambda(t)$ indicate a sequence of tone graphs to specify the underlying lexical tone structures. An F_0 contour on the logarithmic scale of fundamental frequency, as suggested in Fujisaki (1983), is expressed as a scale transformation from $\Lambda(t)$ to $F_0(t)$, corresponding to the syllabic tones fitting themselves with sentence intonation within the vocal range:

$$\frac{\ln F_0(t) - \ln f_{0_b}}{\ln f_{0_t} - \ln f_{0_b}} = \frac{A(\Lambda(t), \zeta(t)) - A(\lambda_b, \zeta(t))}{A(\lambda_t, \zeta(t)) - A(\lambda_b, \zeta(t))}, \quad \text{for } t \geqslant 0, \tag{1}$$

where

$$A(\lambda, \zeta) = \frac{1}{\sqrt{(1-(1-2\zeta^2)\lambda)^2 + 4\zeta^2(1-2\zeta^2)\lambda}}, \quad \text{for } \lambda \geqslant 1, \tag{2}$$

where λ indicates the *square frequency ratio* of a forced vibrating system, and $\zeta \in (0, 0.707)$ the *damping ratio* of the system. Parameters λ_t and λ_b denote respective values of mapping f_{0_t} and f_{0_b} on to a so-called λ-scale ($\lambda \geqslant 1$). Commonly, λ_t and λ_b can be fixed at 1 and 2, respectively. Additionally, $A(\lambda, \zeta)$ indicates *amplitude amplifying coefficients* of a forced vibrating system, also known as *resonance curves* (Fry, 1979). Figure 2(a) shows two resonance curves controlled by both parameters λ and ζ. Since the symmetry $A(\lambda, \zeta) = A(2-\lambda, \zeta)$, the right arm of a resonance curve ($\lambda \geqslant 1$) is basically employed here.

Equations (1) and (2) jointly indicate a structural formulation of the hidden process of articulating syllabic tones and sentence intonation into a final F_0 contour. In Mandarin, syllabic tones are available in a lexicon; each syllable has a tone, and sentence intonation may vary from one utterance to another. Equation (1) states that an F_0 contour $F_0(t)$ is a transformation of a sequence of tone graphs $\Lambda(t)$ modulated by $\zeta(t)$, called a *latent scale*, a hypothetical construct that is invoked to explain observed covariation in tonal behavior. Equation (2) expresses how to perform the transformations when given the latent scale. Technically, this latent scale involved in a transformation tunes the heights and ranges of the tone graphs in F_0, thus characterizing intonation components. This is essentially based on an understanding of the intonation phenomena that, in tonal languages, it is the scale of an F_0 contour that has the intonational function (Abe, 1980); in nontonal ones, it is the shape of the contours. Suppose that these tone graphs were drawn on an elastic transparent sheet, as illustrated by Chao (1968). Stretching the sheet vertically or letting it shrink would alter the heights

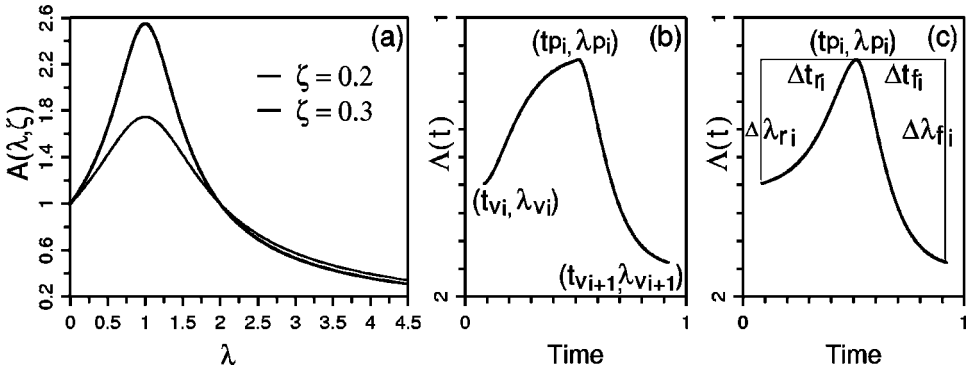


FIG. 2. Examples of: (a) Resonance curves and parametric bell-shaped patterns with both (b) causal and (c) noncausal (rising) transitions.

and ranges of tone graphs. A demonstration is available in Sec. III B (in Fig. 4).

Any fundamental frequency $f_0 \in [f_{0_b}, f_{0_t}]$ is thus controlled by two parameters λ and ζ, corresponding to the effect of syllabic tones and that from factors beyond syllabic tones on F_0. Given any two of f_0, λ, and ζ, the other can be determined by Eqs. (1) and (2). For simple citation, hereafter, let their solutions be expressed as follows:

$$f_0 = T_{f_0}(\lambda, \zeta), \text{ computing } f_0, \text{ given both } \lambda \text{ and } \zeta \text{ values}, \tag{3}$$

$$\lambda = T_{\lambda}(f_0, \zeta), \text{ computing } \lambda, \text{ given both } f_0 \text{ and } \zeta \text{ values}, \tag{4}$$

$$\zeta = T_{\zeta}(\lambda, f_0), \text{ computing } \zeta, \text{ given both } \lambda \text{ and } f_0 \text{ values}. \tag{5}$$

Equation (3) can be directly calculated using Eqs. (1) and (2), while Eqs. (4) and (5) are estimated by an iterative algorithm. A solution to Eq. (5), for example, can be reached by searching ζ from 0.001 to 0.707 in small increments (e.g., 0.001) until ζ satisfies the following conditions:

$$\ln\frac{f_0}{f_{0_b}} > \frac{A(\lambda, \zeta-\delta) - A(\lambda_b, \zeta-\delta)}{A(\lambda_t, \zeta-\delta) - A(\lambda_b, \zeta-\delta)} \times \ln\frac{f_{0_t}}{f_{0_b}}, \text{ and}$$

$$\ln\frac{f_0}{f_{0_b}} < \frac{A(\lambda, \zeta+\delta) - A(\lambda_b, \zeta+\delta)}{A(\lambda_t, \zeta+\delta) - A(\lambda_b, \zeta+\delta)} \times \ln\frac{f_{0_t}}{f_{0_b}},$$

where δ indicates a perturbation (e.g., δ=0.0005), given f_0, λ, f_{0_t}, f_{0_b}, λ_t, and λ_b.

2. Local control

The specification of tone graphs $\Lambda(t)$ consists of placing a series of targets on the time axis. These targets are specified with normalized F_0 values in λ that are always reached in this model. The transitions between two adjacent targets are approximated by second-order filters in such a way that truncated responses connect with minimum discontinuities in their derivatives.

Referring to Fig. 1, a series of targets to specify tone graphs basically consist of tonal valleys and peaks in an alternating fashion. If two peaks occur in succession, such as Tone 1, it is necessary to insert a valley between them and fix its value, for example, at the lower peak; thus, the lower peak takes both peak and valley types. Let (t_{p_i}, λ_{p_i}) indicate the *i*th peak target in the λ-time space, and (t_{v_i}, λ_{v_i}) and $(t_{v_{i+1}}, \lambda_{v_{i+1}})$ denote the valley targets on the left and right sides of the *i*th peak target, respectively. Thus, the *i*th rising and falling (transitions between targets), denoted by $\Lambda_{r_i}(t)$ and $\Lambda_{f_i}(t)$ as functions of time *t*, form a bell-shaped pattern, as sketched in Fig. 2(b).

$$\Lambda_{r_i}(t) = \begin{cases} 2 - ((2-\lambda_{v_i}) + \Delta\lambda_{r_i}(1 - D_{r_i}(t - t_{v_i}))), \\ \quad \text{for } t_{v_i} \leqslant t < t_{p_i}, \\ 0, \quad \text{otherwise}, \end{cases} \tag{6}$$

$$\Lambda_{f_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{f_i}(1 - D_{f_i}(t - t_{p_i})), & \text{for } t_{p_i} \leqslant t < t_{p_{i+1}}, \\ 0, \quad \text{otherwise}, \end{cases} \tag{7}$$

where

$$D_{x_i}(t) = \left(1 + \frac{4.8t}{\Delta t_{x_i}}\right) e^{-\frac{4.8t}{\Delta t_{x_i}}}, \quad x \in \{r, f\}. \tag{8}$$

The transition time for the *i*th rising

$$\Delta t_{r_i} = (t_{p_i} - t_{v_i})/0.985;$$

the transition amplitude for the *i*th rising

$$\Delta\lambda_{r_i} = (\lambda_{v_i} - \lambda_{p_i})/0.95;$$

the transition time for the *i*th falling

$$\Delta t_{f_i} = (t_{v_{i+1}} - t_{p_i})/0.985;$$

and the transition amplitude for the *i*th falling

$$\Delta\lambda_{f_i} = (\lambda_{v_{i+1}} - \lambda_{p_i})/0.95. \tag{9}$$

In Eq. (6), $(2-\lambda_{v_i})$ expresses a conversion of λ_{v_i} from $\lambda \geqslant 1$ to $\lambda \leqslant 1$ (the left arm of a resonance curve) by using the symmetry $A(2-\lambda, \zeta) = A(\lambda, \zeta)$. Accordingly, both rising and falling transitions are calculated using an identical causal transition mechanism. The resulting rising transitions are further converted back to $\lambda \geqslant 1$ in order to share the same parameter setup, $\lambda_b = 2$, with the falling transitions in the phase of computing $F_0(t)$. Since the transition expressed in Eq. (6) is causal, it is henceforth referred to as a *causal transition.*

TABLE I. Citation values of tonal targets in the experiments with three speakers.

Tone k	Tone 1		Tone 2			Tone 3			Tone 4			
P_{kj}/V_{kj}	P11	P12	P21	V21	P22	P31	V31	P32	P41	V41	$V41^h$	$V41^*$
$f_{c_{kj}}$	$f_{c_{11}}$	$f_{c_{12}}$	$f_{c_{21}}$	$f_{c_{22}}$	$f_{c_{23}}$	$f_{c_{31}}$	$f_{c_{32}}$	$f_{c_{33}}$	$f_{c_{41}}$	$f_{c_{42}}$	$f^h_{c_{42}}$	$f^*_{c_{42}}$
FY	265	263	n/a	171	243	185	151	180	187	148	167	158
FL	289	284	212	204	282	196	157	208	302	151	226	188
WL	370	363	n/a	250	355	235	200	280	375	176	275	225

Also, there is a *noncausal transition* for approximating the *i*th rising:

$$\Lambda_{r_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{r_i}(1 - D_{r_i}(t_{p_i} - t)), & \text{for } t_{p_{i-1}} \leqslant t \leqslant t_{p_i}, \\ 0, & \text{otherwise}. \end{cases} \tag{10}$$

Figure 2(c) illustrates corresponding bell-shaped patterns with the noncausal transitions. In spite of slight differences in transition profiles, both valley and peak targets are consistently reached by either transition with the parameter conversion in Eq. (9), where constant coefficients 0.95 and 0.985 particularly depend on the constant coefficient 4.8 fixed in Eq. (8). One relative benefit, in comparison with the causal transitions, is that tonal targets (including peaks and valleys) can be completely specified by transition time Δt_{x_i} and amplitude $\Delta\lambda_{x_i}$, $x \in \{r, f\}$, with an initial target, thus facilitating the parameter estimation with a data-driven fitting process.

$\Lambda(t)$ is eventually expressed as a concatenation of n truncated bell-shaped patterns lined up in series on the time axis:

$$\Lambda(t) = \Lambda_{r_1}(t) + \sum_{i=1}^{n-1} \min(\Lambda_{f_i}(t), \Lambda_{r_{i+1}}(t)) + \Lambda_{f_n}(t), \tag{11}$$

where n indicates the number of tonal peaks involved in the lexical tone structures, and assume $t_{p_{n+1}} = \infty$. Note that the noncausal transition, hereafter, is employed to approximate the rising transitions.

The model has proven effective for obtaining appropriate approximations of, for example, 968 observed F_0 contours from eight native speakers with a small number of parameters where $\zeta(t)$ was fixed at a default value. These parameters are localized and thus suited to a data-driven fitting process, as exemplified in Ni and Kawai (2003).

III. DESCRIPTION OF THE APPROACH

Drawing upon the theory of tone and intonation (Chao, 1968; Xu, 1997; Kratochvil, 1998), we assume that the underlying lexical tone structures of observed F_0 contours provide a construct for coupling both syllabic tones and sentence intonation together with specific temporal structures. To separate the intonation components (sentence intonation) from an observed F_0 contour, first we need a *norm* (or *citation form*) to specify the lexical tone structures. Second, we need a technique for estimating tonal variations from the norm to the observed F_0 contour. As a consequence, we make a number of specific proposals for implementing this approach on a computer.

A. Defining citation values for tonal targets

The time-frequency curves of lexical tones produced in isolation are rather stable (Shen, 1985, 1990) and are known as *citation forms* of lexical tones. Since there theoretically exist only four citation forms of lexical tones in a voice, we can then define *citation values* for tonal targets, focusing particularly on the targets that sparsely specify the citation forms of lexical tones. Consequently, the citation values of tonal targets specify the relative relationships among the four lexical tones in the voice and neutral mood.

In practice, the citation values of tonal targets for a speaker can be determined by averaging several isolated syllable samples produced by the speaker. Table I lists the citation values of tonal targets for three native speakers used in the evaluation experiments described in Sec. IV. More specifically, let $f_{c_{kj}}$ indicate citation values for the *j*th target of Tone k, $k=1$, 2, 3, and 4; $j \in \{1,2,3\}$. Each tonal target has a citation value except for Tone 4. The second tonal target of Tone 4 has three typical variants: Citation value $f_{c_{42}}$ obtained by averaging several isolated samples, $f^h_{c_{42}}$ for half-Tone 4, and $f^*_{c_{42}}$ for Tone 4 followed by any tone except for Tone 4. When Tone 4 is followed by another Tone 4, it appears as half-Tone 4 (Chao, 1968); when it occurs at the end of an utterance even in isolation, the downdrift phenomenon usually occurs. Technically, we simply approximate $f^h_{c_{42}} = (f_{c_{41}} + f_{c_{42}})/2$, and $f^*_{c_{42}} = (f_{c_{42}} + f^h_{c_{42}})/2$. As for Tone 0, it will simply copy the citation value of the last target of the preceding tone.

B. Tone transformation

Transformations are the essence of a generative theory. A tone transformation shall specify a set of procedures which, applied to the elements of a structure of tone, will produce all the more complex forms from their citation values. The term *tone transformation* here means a computational method for this purpose.

By using Eq. (4), each citation value $f_{c_{kj}}$ can be converted into a λ value, $T_\lambda(f_{c_{kj}}, \zeta_0)$, with respect to ζ_0 (a reference value). On the other hand, $T_\lambda(f_{c_{kj}}, \zeta_0)$ can be used to generate numerous F_0 values, denoted by $f_{0_{kj}}$, when ζ alters from 0 to 0.707. Thus, a constrained tone transformation can be expressed as

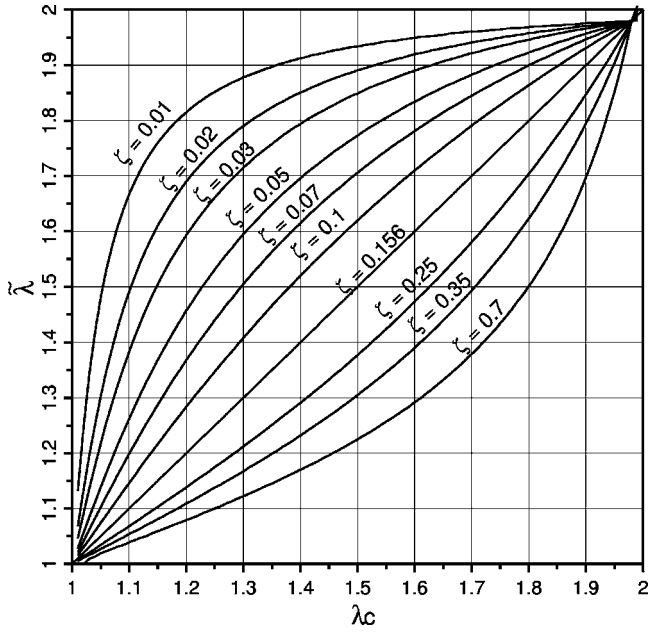


FIG. 3. Exhibition of the theoretical constraints underlying the tone transformations.

$$f_{0_{kj}} = T_{f_0}(T_\lambda(f_{c_{kj}}, \zeta_0), \zeta), \quad k = 1,2,3,4;\ j \in \{1,2,3\}. \quad (12)$$

Hereafter, assume $\zeta_0 = 0.156$.

Let $\lambda_c = T_\lambda(f_{c_{kj}}, \zeta_0)$ and $\tilde{\lambda} = T_\lambda(f_{0_{kj}}, \zeta_0)$. To examine the constraints involved in the tone transformation, we rewrite Eq. (12) into the following form:

$$\tilde{\lambda} = T_\lambda(T_{f_0}(\lambda_c, \zeta_0), \zeta). \quad (13)$$

Figure 3 exhibits the theoretical relationships between both λ_c and $\tilde{\lambda}$ in $[1,2]$ at several ζ values. Basically, when ζ departs from the reference value ζ_0(=0.156), $\lambda_c \in [1,2]$ is nonlinearly transformed to $\tilde{\lambda}$, whose span is sharply narrowed near both ends of range $[1,2]$.

To balance ζ on both sides of ζ_0, a normalized damping ratio ζ_n is defined as follows.

$$\zeta_n = \begin{cases} (\zeta - \zeta_0)/(0.7 - \zeta_0), & \text{for } \zeta \geqslant \zeta_0, \\ (\zeta - \zeta_0)/\zeta_0, & \text{for } \zeta < \zeta_0, \end{cases} \quad (14)$$

Obviously, $\zeta_n \in [-1,1]$ can be converted back to ζ, denoted by $\tilde{\zeta}$:

$$\tilde{\zeta} = \begin{cases} (0.7 - \zeta_0) \times \zeta_n + \zeta_0, & \text{for } \zeta_n \geqslant 0, \\ \zeta_0 \times \zeta_n + \zeta_0, & \text{for } \zeta_n < 0. \end{cases} \quad (15)$$

Figure 4 demonstrates the tone transformation applied to the citation forms of lexical tones by speaker WL. Her vocal range is approximated by [100 Hz, 500 Hz], a slight exten-

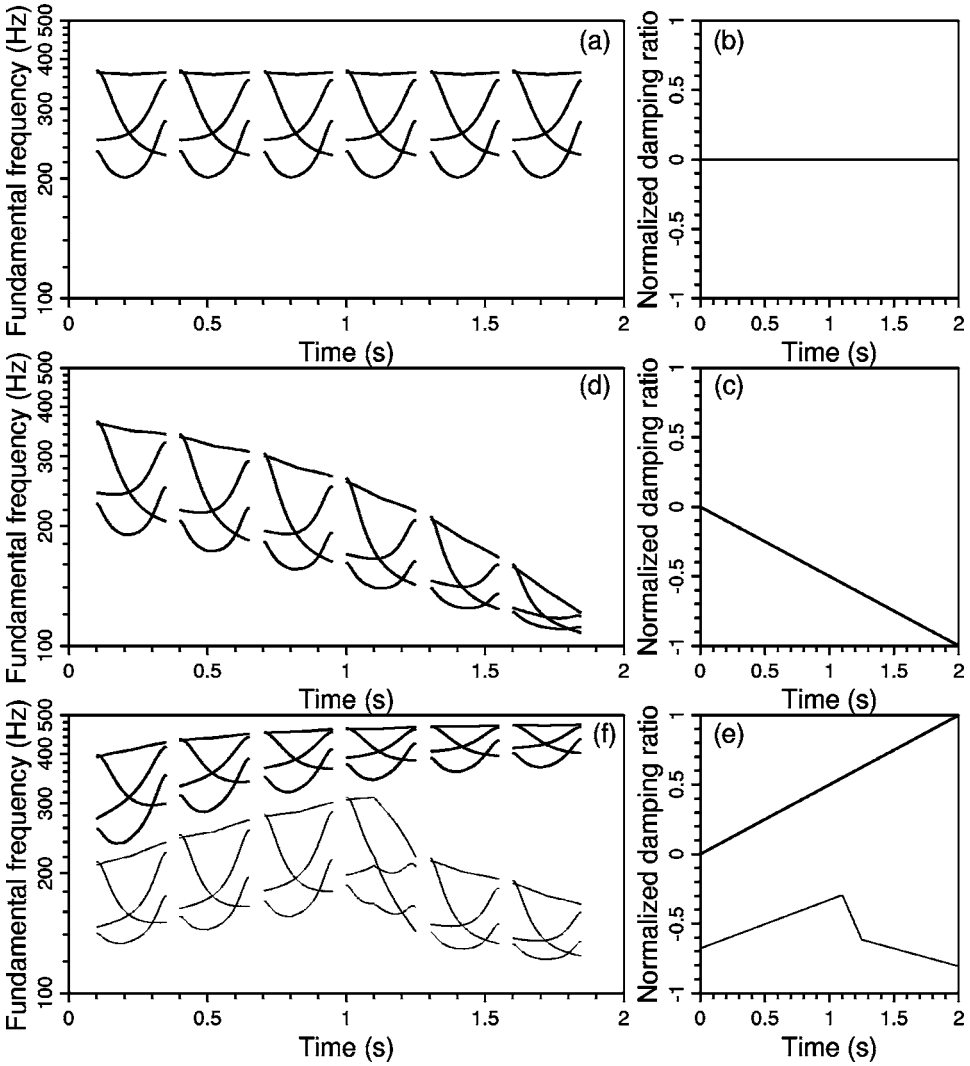


FIG. 4. Demonstration of tone transformations from the citation forms in (a) with $\zeta(t) = \zeta_0$ shown in (b) to the transformation forms in (d) and (e) with varying $\zeta(t)$ displayed in (c) and (e), respectively.

sion of the real vocal range. Panel (a) shows the citation forms of four lexical tones, which are the output of an algorithm (see Sec. III D 2) to the input of sequences of citation values of tonal targets taken from Table I with an assumed temporal structure. The four citation forms are superimposed together over time and are repeated six times. Panel (b) shows $\zeta_n(t)=0$, corresponding to $\zeta(t)=\zeta_0$. When $\zeta_n(t)$ changes linearly from 0 to −1 during a two-second interval, as shown in (c), these citation forms in (a) will change into the transformation forms shown in (d). While $\zeta_n(t)$ appears as the thick lines shown in (e), corresponding transformation forms are indicated by the thick lines shown in (f). Technically $\zeta(t)$ anchors these tone graphs $\Lambda(t)$in the vocal range, and simultaneously modulates their heights and ranges in F_0. The F_0 ranges of tone graphs are narrowed at the extra high/low regions of a vocal range. Actually, a typical feature of low pitch is the narrowness of pitch ranges (Chao, 1968).

This demonstration shows that it is possible to transform the citation forms of four lexical tones to possible variants in the vocal range. The relationships between citation and transformation forms of the four lexical tones may be compared to printed and handwritten forms of Chinese characters, where the internal structures of characters remain unchanged in both forms in that humans can still recognize the characters. For lexical tones to help distinguishing word meaning, it is definitely important for a lexical tone to distinguish itself from the others regardless of sentence intonation.

C. Estimating intonation components

A question may arise as to why the model enables the tone transformation technique to separate tone and intonation components from the acoustic correlates (F_0 contours). From the viewpoint of speech production, the F_0 control mechanisms dealt with by the motor system offer multiple solutions to accomplish similar tasks (Titze, 1997). In practice, as exemplified in Ni and Hirose (2000), multiple solutions of $\Delta t_x, \Delta\lambda_x$, and $\zeta(t)$ exist for a set of observed F_0 contours in the sense of generating very closed approximations of them, given the lexical tone structures. When values of parameters Δt_x and $\Delta\lambda_x$ for syllabic tones take those suited for their citation forms, corresponding $\zeta(t)$ values would obviously characterize *tonal variation* from the citation to the observed values. Needless to say, a syllabic tone specified by its citation form also has the function of distinguishing word meaning.

1. Measuring tonal variation

Let $\hat{f}_{0_{kj}}$ denote observed values for the jth target of Tone k, $k=1,2,3$, and 4; $j\in\{1,2,3\}$ from real speech (hereafter, a variable with ^ takes observed values). Thus $\hat{f}_{0_{kj}}$ can be regarded as a transformation of corresponding citation value $f_{c_{kj}}$. Then, we have

$$\zeta_{kj}=T_\zeta(T_\lambda(f_{c_{kj}},\zeta_0),\hat{f}_{0_{kj}}), \tag{16}$$

where ζ_{kj} indicates tonal variation from citation values $f_{c_{kj}}$ to observed values $\hat{f}_{0_{kj}}$.

Technically, the internal variation among the four lexical tones is captured by sets of parameter values $T_\lambda(f_{c_{kj}},\zeta_0)$. Accordingly, differences among the citation forms of tones become transparent to measuring tonal variation, thus enabling ζ_{kj} to closely characterize the intonation components. When $\zeta_{kj}>\zeta_0$, observed F_0 values of a tonal target are larger than its citation value. Similarly, when $\zeta_{kj}<\zeta_0$, observed F_0 values are less than the corresponding citation value. Regarding a target, the larger the $\hat{f}_{0_{kj}}$, the larger the ζ_{kj}.

2. Derivation of sparser specifications

Given an observed F_0 contour, a sparser specification of it would require derivation of a sequence of tonal targets, taking into account its segmental labeling, such as syllable boundary and *pinyin* transcription. This derivation may involve two logic steps:

- Determine tonal F_0 peaks, and
- Determine tonal valleys for each section of F_0 contours partitioned by these F_0 peaks, taking into account the underlying lexical tone structures.

Ni and Kawai (2003) describes an effort made to extract tonal F_0 peaks and gliding features from observed F_0 contours with the model. An algorithm there works on an F_0 contour with such input information as segmental labeling in two substeps. First, tonal F_0 peaks are determined by an analysis-by-synthesis-based pattern-matching technique. More specifically, several baseline tone templates obtained by clustering techniques are prestored for each tone type in a parametric form. Every tone template applied to a syllabic tone is adjusted so as to fit its parameters, with the part of F_0 contours coinciding with the tone-carrying syllable as accurately as possible. The number of tonal F_0 peaks is then ascertained by finding a path with the minimum matching errors among all of the possible paths within the lattice resulting from the process of multiple-tone-template fitting, taking into account the underlying lexical tone structures. Second, tone gliding features, characterized by parameters Δt_{x_i} and $\Delta\lambda_{x_i}$, $x\in\{r,f\}$, are re-estimated with reference to the tonal F_0 peaks by minimizing the errors between the model-approximated and measured contours. Experimental results indicated that 94% of the automatically estimated tonal F_0 peaks were consistent with the manual labels in an open test of 968 utterances from eight native speakers.

Tonal valleys are then determined by analyzing the resynthesized contours $\Lambda(t)$ obtained by the algorithm at the phase of determining tonal peaks. A tonal valley is assumed for every syllabic tone except for Tone 1, which needs no tonal valley. Let (t_{v_i},λ_{v_i}) and $(t_{v_{i+1}},\lambda_{v_{i+1}})$ be two assumed valleys on the left and right sides of the ith peak. Initiate the valley $(t_{v_{i+1}},\lambda_{v_{i+1}})$, for example, at the lowest of $\Lambda(t)$ between the ith peak (t_{p_i},λ_{p_i}) and the next, denoted by $(t_{v_{min_{i+1}}},\lambda_{v_{min_{i+1}}})$. Then, search through it along with $\Lambda(t)$ toward the ith peak by decreasing $t_{v_{i+1}}$ in short steps (e.g., 0.005 s) until reaching the following conditions:

$$\lambda_{v_{i+1}} - \lambda_{p_i} \leqslant (\lambda_{v_{\min_{i+1}}} - \lambda_{p_i}) \times 0.95, \text{ or } t_{v_{i+1}} = t_{p_i}.$$

If distances between two consecutive estimated valleys are less than a threshold, such as 0.025 s, only a valley is then assumed at the midpoint between them.

3. Separation algorithm

An algorithm is presented below for estimating intonation components, given the lexical tone structures and a sparser specification of the F_0 contours, $(\hat{t}_i, \hat{f}_{0_i})$, $i=1,\ldots,N$, where N is the number of observed targets; $\hat{t}_i$ and $\hat{f}_{0_i}$ indicate the timing and F_0 value of the ith target, respectively. These targets are intended to specify the prosodic contributions of syllabic tones and sentence intonation to the F_0 contours. This estimation, then, consists of transforming target values in pairs from the citation to the observed values, which are aligned in time in terms of the underlying lexical tone structures.

Input:

- The sparser specifications: $(\hat{t}_i, \hat{f}_{0_i})$, $i=1,\ldots,N$,
- the underlying lexical tone structures,
- the citation values of tonal targets: $f_{c_{kj}}$, $k=1,2,3$, and 4; $j \in \{1,2,3\}$, and
- the vocal range $[f_{0_b}, f_{0_t}]$.

Step 1: Associate observed targets with the elements of lexical tone structures by rules:

Rule 1: A peak is only associated with a syllabic tone.
Rule 2: A valley may be associated with one or two syllabic tones.
Rule 3: No cross-association is allowed.

Step 2: Determine F_0 pairs $(f_{c_{kj}}, \hat{f}_{0_{kj}})$ taking into account the following rules:

- If a valley is outside the scope of a syllable, delete the F_0 pair associated with the tone carried by the syllable; and
- $f_{c_{0_j}}$ for Tone 0 takes the citation value of the last target of its preceding tone.

Step 3:

- Compute $\zeta_{kj} = T_\zeta(T_\lambda(f_{c_{kj}}, \zeta_0), \hat{f}_{0_{kj}})$ for each tone including Tone 0, $j \in \{1,2,3\}$, and
- Convert these ζ_{kj} values to corresponding ζ_n (normalized values) using Eq. (14).

Output: $(\hat{t}_i, \zeta_{n_i})$, $i=1,\ldots,N$, characterizing the intonation components.

Three rules are used at Step 1 to guide the association of these observed targets, $(\hat{t}_i, \hat{f}_{0_i})$, $i=1,\ldots,N$, with the elements of the underlying lexical tone structures, thus giving $\hat{f}_{0_{kj}}$ used at Step 2. A peak inside a syllable is only associated with the syllabic tone, while a valley may be associated with one or two tones. Rule 1 reflects the assumption that the tone domain is within a syllable (Lin, 1995, and others). Since Mandarin tones have to perform the function of distinguishing word meaning, they are rather similar to segmental features like a phoneme in a word. Rule 2 reflects the aspect of suprasegmental features in that lexical tone structures also construct a framework to carry sentence intonation. Rule 3 is general in an autosegmental theory.

D. Combination algorithm

As an inverse process of the separation, it is possible to generate F_0 contours by superimposing given intonation components on the lexical tone structures whose elements are specified by the citation values of tonal targets, given the temporal structures. An algorithm for combining tone and intonation would involve two operations:

- Compute a series of targets by superimposing the intonation components on the citation forms of lexical tone structures, and
- Compute the F_0 contours by connecting the transitions through these targets.

1. Computing target F_0 values

An algorithm is presented below for computing target F_0 values, given the prosody specifications and temporal structures.

Input:

- The citation values of tonal targets, $f_{c_{kj}}$, $k=1,2,3,4$; $j \in \{1,2,3\}$,
- an intonation component $\zeta_n(t)$,
- the lexical tone structures,
- the intended temporal structures: the timing of targets, t_i, $i=1,\ldots,N$, and
- the vocal range $[f_{0_b}, f_{0_t}]$.

Step 1: Compute $\zeta_{n_i} = \zeta_n(t_i)$, and convert ζ_{n_i} to $\tilde{\zeta}_i$, $i=1,\ldots,N$, using Eq. (15).

Step 2: Assign citation values $f_{c_{kj}}$ to the elements of lexical tone structures:

- In the case of Tone 4, if it is followed by another Tone 4, $f^h_{c_{42}}$ replaces $f_{c_{42}}$. Otherwise, if it is followed by any tone other than Tone 4, use $f^*_{c_{42}}$ to replace $f_{c_{42}}$.
- As for Tone 0 ($k=0$), citation value $f_{c_{0j}}$ takes the last one of its preceding tones. Let f_{c_i}, $i=1,\ldots,N$, indicate the sequence of citation values that specify the elements of lexical tone structures.

Step 3: Compute target F_0 values (predicted values) $f_{0_i} = T_{f_0}(T_\lambda(f_{c_i}, \zeta_0), \tilde{\zeta}_i), i=1,\ldots,N$.

Output: (t_i, f_{0_i}), $i=1,\ldots,N$, as a sparser specification of the prosodic contributions.

2. Computing F_0 contours

An algorithm for computing contours by the model fed with a sparser specification of the prosodic contributions involves the following steps.

Input:

- A sparser specification; let (t_j, f_{0_j}) be the jth target, $j=1,\ldots,N$,
- the vocal range $[f_{0_b}, f_{0_t}]$.

Step 1:

- Compute $\lambda_j = T_\lambda(f_{0_j}, \zeta_0)$ for all of the target F_0 values,
- take out all of the n peaks (t_{p_i}, λ_{p_i}), $i=1,\ldots,n$,
- let (t_{v_i}, λ_{v_i}) denote the valley just before the ith peak, and
- let $(t_{v_{i+1}}, \lambda_{v_{i+1}})$ denote the valley just after the ith peak.

Step 2: Compute the ith transition time and amplitude, $i=1,\ldots,n$:

- transition time for the ith rising: $\Delta t_{r_i} = \max(0.05, (t_{p_i} - t_{v_i})/0.985)$,
- transition amplitude for the ith rising: $\Delta\lambda_{r_i} = \max(0.02, (\lambda_{v_i} - \lambda_{p_i})/0.95)$,
- transition time for the ith falling: $\Delta t_{f_i} = \max(0.05, (t_{v_{i+1}} - t_{p_i})/0.985)$, and
- transition amplitude for the ith falling: $\Delta\lambda_{f_i} = \max(0.02, (\lambda_{v_{i+1}} - \lambda_{p_i})/0.95)$.

Step 3: Compute the tone graphs $\Lambda(t)$ using Eq. (11) with these parameter values.

Step 4: Compute $F_0(t) = T_{f_0}(\Lambda(t), \zeta_0)$ using Eq. (3).

Output: $F_0(t)$ as the synthetic F_0 contours.

If two consecutive peaks occur in the sparser specification, it is necessary to assume a valley between them and fix its value, for example, at the lower peak. In addition, default valleys are assumed before the first peak and after the last peak if no valley exists there.

IV. EXPERIMENTAL CONFIRMATION AND DISCUSSIONS

The validity of an approach can be tested by its ability to analyze observed samples. For this purpose, we select a number of speech samples from existing Chinese speech corpora, such as USTC95 (Wang *et al.*, 1996). This evaluation is conducted in three experiments. Experiment 1 demonstrates this approach with an example. Through Experiment 2, we test the reliability of this approach for estimating intonation components and show the extensibility of the tone transformation technique applied to estimate intonation variation. With Experiment 3, we justify the two assumptions mentioned in Sec. II A.

A. Experiment 1: Demonstration of the approach

This demonstration includes: (1) Sparser specifications of F_0 contours by a series of targets; (2) approximations of F_0 contours by connecting transitions through these targets; (3) association of these targets with the elements of underlying lexical tone structures; (4) extraction of intonation components by the tone transformation technique; and (5) resynthesis of F_0 contours by superimposing intonation components on the lexical tone structures whose elements are specified by citation values of tonal targets. This is done by analysis of an observed F_0 contour ["+" sequences shown in Fig. 5(a)] for the sentence *zhei4li3you3ren2lai2*, *bu4neng2yi2* [someone may come here, (it) cannot move away.] uttered by speaker FY. Her vocal range $[f_{0_b}, f_{0_t}]$ is fixed at [120 Hz, 420 Hz], and the citation values of tonal targets by FY are presented in Table I.

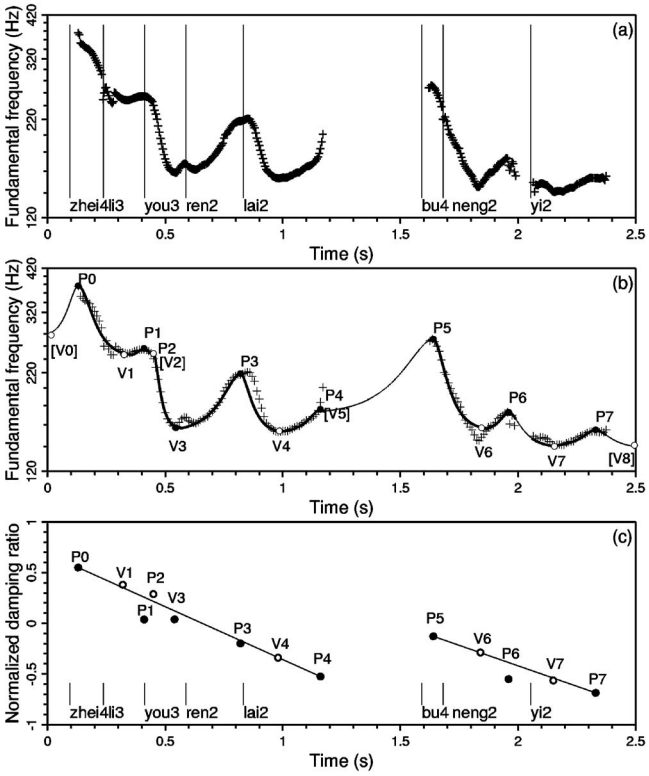


FIG. 5. Observed F_0 ["+," every 5 ms in (a), and every 10 ms in (b)] for a sentence *zhei4li3you3ren2lai2*, *bu4neng2yi2* (Someone may come here, (it) cannot move away). The vertical lines in (a) indicate the beginning of syllables. The continuous lines in (b) show an approximation of the F_0 contour given by transitions through targets P0–P7 and V0–V8. The intonation components extracted from these targets are displayed in (c).

The prosodic contribution of tone and intonation to the F_0 contour is sparsely specified by the following 13 targets [white and black circles shown in Fig. 5(b)]:

$$\langle \mathrm{P0}, \mathrm{V1}, \mathrm{P1}, \mathrm{P2}, \mathrm{V3}, \mathrm{P3}, \mathrm{V4}, \mathrm{P4}, \mathrm{P5}, \mathrm{V6}, \mathrm{P6}, \mathrm{V7}, \mathrm{P7} \rangle.$$

These observed target values $(\hat{t}_i, \hat{f}_{0_i})$, $i=1,\ldots,13$, can be found in Table II.

The continuous lines shown in Fig. 5(b) display the transitions through these targets. Table II lists values of model parameters converted from these targets by Eq. (9). [V2] and [V5] are two insertion valleys to separate consecutive peaks. [V0] and [V8] are two default valleys assumed before the first peak and after the last peak. The rooted-mean-square errors (RMSEs) between observed and re-synthesized F_0 contours are 0.038 (around 4.48 Hz in absolute errors). This good approximation by the transitions through these targets indicates that it is these targets that capture the main characteristics of the F_0 contour, while the exact shape of the F_0 contour can be predicted by the model, for example, given these targets.

Table III lists the association of these targets (the second row) with the underlying lexical tone structure (the first row), where Tone 3 of *you3* takes its half-form. Valley target

TABLE II. Values of the model parameters converted from a series of targets.

i	Peaks $(\hat{f}_{0_i})$	$(\hat{t}_{p_i},\hat{\lambda}_{p_i})$	Valleys $(\hat{f}_{0_i})$	$(\hat{t}_{v_i},\hat{\lambda}_{v_i})$	$\Delta t_{r_i},\Delta\lambda_{r_i}$	$\Delta t_{f_i},\Delta\lambda_{f_i}$
1	P0 (377)	(0.13,1.12)	[V0] (275)	(0.01,1.27)	0.12, 0.16	0.19, 0.22
2	P1 (256)	(0.41,1.30)	V1 (243)	(0.32,1.33)	0.09, 0.03	0.05, 0.02
3	P2 (248)	(0.45,1.32)	[V2](248)	(0.45,1.32)	0.05, 0.02	0.09, 0.33
4	P3 (219)	(0.82,1.39)	V3 (156)	(0.54,1.63)	0.28, 0.25	0.16, 0.28
5	P4 (176)	(1.16,1.53)	V4 (151)	(0.98,1.66)	0.18, 0.14	0.05, 0.02
6	P5 (271)	(1.64,1.28)	[V5] (176)	(1.16,1.53)	0.48, 0.26	0.20, 0.38
7	P6 (173)	(1.96,1.54)	V6 (154)	(1.84,1.64)	0.12, 0.11	0.19, 0.22
8	P7 (155)	(2.33,1.64)	V7 (139)	(2.15,1.75)	0.18, 0.12	0.16, 0.13
9			[V8] (138)	(2.49,1.76)		

V1, for instance, is associated with both Tones 2 and 4, while V1 is actually carried by syllable *li3* whose surface tone is Tone 2 according to tone sandhi rules. Accordingly, 16 targets are used, two for each tone in accordance with the configurations shown in Fig. 1.

Figure 5(c) shows the intonation components extracted from these targets (circles). Table IV lists a part of the parameter values that occurred in the process of estimating the intonation components by the separation algorithm (Sec. III C 3). The estimate values, ζ_{n_i}, $i=1,\ldots,13$, approximately wave along with decline lines P0P4 and P5P7, which can be expressed as

$$\zeta_n(t)=\begin{cases}-1.045t+0.686, & \text{for } 0\leqslant t<1.5,\\ -0.809t+1.198, & \text{for } t\geqslant 1.5.\end{cases} \tag{17}$$

It is noticeable from this example that, while the observed target values, indicated by white and black circles in Fig. 5(b), undulate considerably, the remains (ζ_n values) undulate much less after suppressing the effects of lexical tones on the F_0 contour by the tone transformation technique. Had syllabic tones (a sequence of primary and nonprimary targets) been compared to the stress group described in Thorsen (1980), the results shown in Fig. 5(c) would have been intonational contours. From this point of view, both syllabic tones and sentence intonation are superimposed together to form the speech melody, but the "superimposition" here is not simple addition.

Figure 6(c) plots the synthesized contours (continuous lines) by connecting the transitions through a series of predicted targets (white and black circles). These targets are synthesized by superimposing intonation components [lines shown in Fig. 6(b)] on the lexical tone structure whose elements are specified by the citation values of tonal targets by FY [white and black triangles shown in Fig. 6(a)]. The intonation components are expressed in Eq. (17), while the lexical tone structure and its temporal structure are adopted from the example shown in Fig. 5. The RMSEs between observed and synthesized F_0 contours are 0.055 (around 7.42 Hz in absolute errors). By comparing the synthesized F_0 contours with the observed F_0 contours ("+" sequences), two portions around $\widetilde{P}1$ and $\widetilde{P}6$ more or less diverge from to each other. This reflects the divergence between the predicted values [$\widetilde{P}1$ and $\widetilde{P}6$ in Fig. 6(b)] and the measured values [P1 and P6 shown in Fig. 5(c)]. This observation indicates that the intonation components should be piecewise linear rather than straight.

TABLE III. Association of a series of targets with underlying lexical tone structures.

Tone 4	Tone 2	Tone 3	Tone 2	Tone 2	Tone 4	Tone 2	Tone 2
P0, V1	V1, P1	P2, V3	V3, P3	V4, P4	P5, V6	V6, P6	V7, P7

B. Experiment 2: Case analysis

The validity of the proposed method shall be evaluated by its ability to extract intonation components from various speech samples in a statistical sense. The difficulty arising in practice is that no true answer of intonation components is known in a strict sense. As Kratochvil (1998) remarked, as far as the specified properties of Chinese intonation are concerned, this is still an area of much speculation. Nevertheless, there have since been several systematic investigations into the relation between tone and intonation with different emphases on, for example, effects of tones and focus (Xu, 1999), interaction of tone and intonation analyzed into a set of tags (parameters) in Stem-ML (Kochanski and Shih, 2003), quantitative decomposition of F_0 contours into (positive and negative) accent and phrase components (Fujisaki, 2004), and statistics-based modeling of the effects from factors other than speakers, tones and contextual tones as "prosodic states" (Chen *et al.*, 2005). That no satisfactory simple answer to "the specified properties of Chinese intonation" has since been forthcoming is in itself an indication of the complexity of this issue. For this reason, we confine our experiments to an analysis of several read speech samples, minimizing the effects from factors other than syllabic tones

TABLE IV. Values of parameters occurring in extracting an intonation component.

i	1	2	3	4	5	6	7	8
Targets	P0	V1	P1	P2	V3	P3	V4	P4
$\hat{t}_i$	0.13	0.32	0.41	0.45	0.54	0.82	0.98	1.16
P_{kj}/V_{kj}	P41	V21	P22	P31	V31	P22	V21	P22
$\hat{f}_{0_i}$ (or $\hat{f}_{kj}$)	377	243	256	248	156	219	151	176
$f_{c_{kj}}$	187	171	243	185	151	243	171	243
$T_\lambda(f_{c_{kj}},\zeta_0)$	1.25	1.54	1.33	1.49	1.66	1.33	1.54	1.33
ζ_{kj}	0.455	0.363	0.176	0.313	0.177	0.125	0.103	0.074
ζ_{n_i}	0.550	0.381	0.037	0.289	0.039	−0.20	−0.34	−0.53

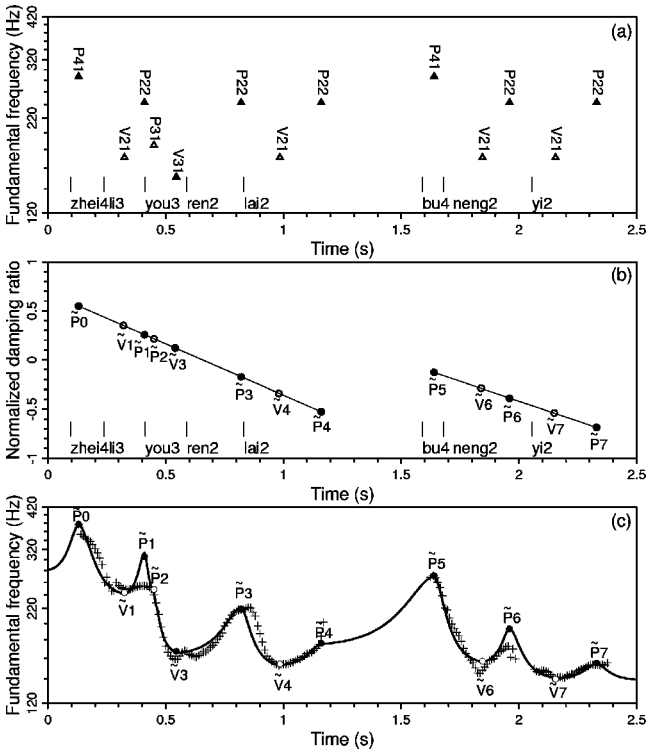


FIG. 6. Example of synthesizing F_0 contours by superimposing an intonation component on the citation forms of a lexical tone structure. Panel (a) displays the citation values of tonal targets used to specify the citation forms of the specific lexical tone structure. Panel (b) plots the assumed intonation components. Panel (c) shows the synthesized contours (continuous lines) through a series of targets (circles), which sparsely specify the prosodic contributions of both the specified lexical tone structure in (a) and intonation components in (b); the "+" sequences indicate the same F_0 contour as shown in Fig. 5.

and sentence intonation. The evaluation method basically is to compare RMSEs between observed and resynthesized F_0 contours as demonstrated in Experiment 1, and explain how the experimental results fit with the current account of Chinese intonation or the findings on intonation in other languages.

Four experiments are conducted. First, we systematically investigate the "intonation components" in isolated words including mono- and trisyllable words that cover various tone combinations. Second, this investigation is extended to limited "designed" short sentences uttered as statements and questions with *ma0* (a question particle). Third, we extract the intonation components involved in conventional Chinese greetings. Finally, we perform statement-question intonation conversion to demonstrate that the measured intonation variation can be reproduced in a perceptual sense. All of these speech samples were read by a female native speaker (FL) except for the greetings by WL and adopted from an existing speech corpus (Wang *et al.*, 1996). FL's vocal range $[f_{0_b}, f_{0_t}]$ is fixed at [100 Hz, 500 Hz], a slight extension of the frequency ranges measured from these speech samples. The citation values of tonal targets by FL are listed in Table I.

1. Isolated words

Seven hundred and twelve isolated words are used for investigating neutral intonation components: 208 isolated syllables (52 distinct base syllables $\times 4$ tones) and 504 trisyllable words, which cover 64 ($=4\times4\times4$) tone combinations; each tone combination distributes to five to nine samples. These isolated syllables are also used for estimating the citation values of tonal targets for the speaker by the method described in Sec. III A.

Figure 7 displays the scattered target F_0 values measured from these isolated syllables in the left panels from (a) to (d) and corresponding ζ_{n_i} values, $i \in \{1,2,3\}$, in the right panels from (e) to (h). In each panel, a symbol (cross, tri-

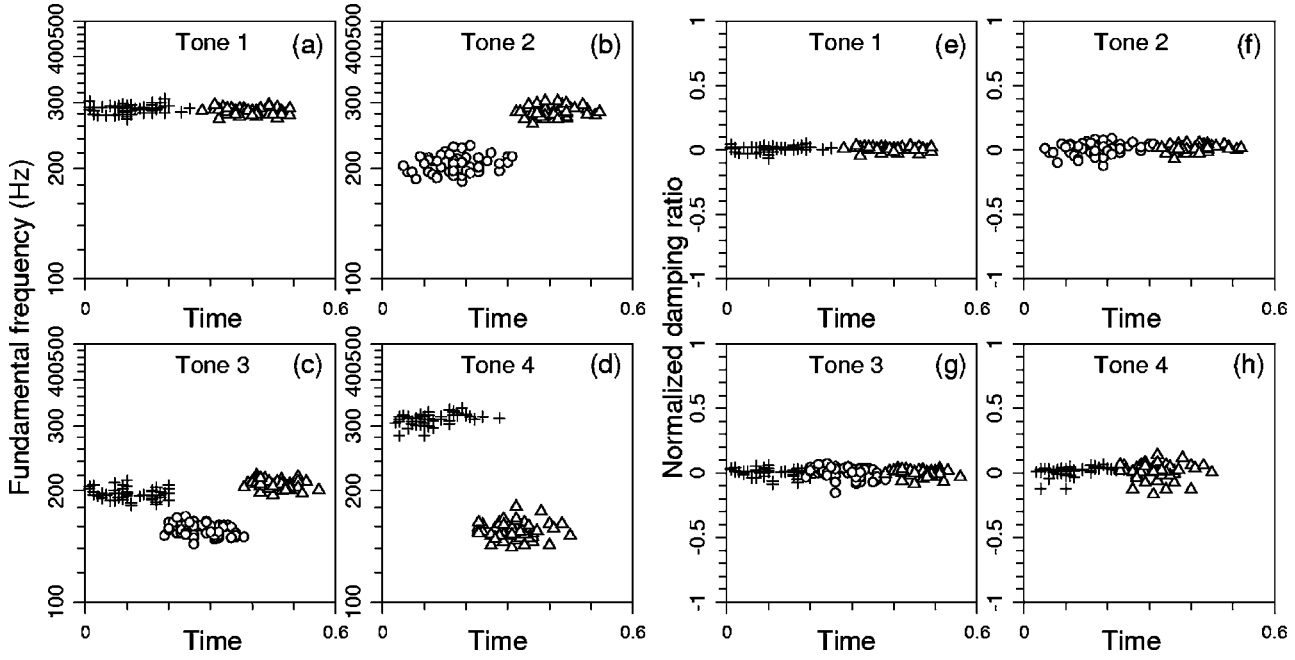


FIG. 7. Scattered F_0 values of tonal targets measured from 208 isolated syllables in the left panels from (a) to (d), and corresponding normalized damping ratios extracted from these targets in the right panels from (e) to (h).

TABLE V. RMSEs between observed and re-synthesized F_0 contours (see text).

	Count	ϵ^i_ζ	$\epsilon^i_{\zeta_n}$	$\epsilon^{m1}_{\zeta_n}$	$\epsilon^{m2}_{\zeta_n}$
Syllables	208	0.0533	0.0570	0.0657	0.0665
Words	504	0.0575	0.0758	0.0828	0.0853
Sentences	48	0.0477	0.0667	0.0767	n/a

angle, or circle) corresponds to a tonal target (e.g., P11 in Fig. 1). The mean of all of these ζ_{n_i} values is 0.0089. The RMSEs between observed and resynthesized F_0 contours are listed in Table V (row "Syllables"). Different RMSEs ε^i_ζ, $\varepsilon^i_{\zeta_n}$, $\varepsilon^{m1}_{\zeta_n}$, and $\varepsilon^{m2}_{\zeta_n}$ indicate which ζ values are input to the combination algorithm (Sec. III D) as the intonation components for re-synthesizing the F_0 contours. Column ε^i_ζ indicates RMSEs with the re-synthesized F_0 contours generated by individual ζ_{kj} values calculated by Eq. (16); $\varepsilon^i_{\zeta_n}$, those by individual normalized ζ_{kj} values converted through Eq. (14); $\varepsilon^{m1}_{\zeta_n}$, those by respective means of these normalized ζ_{kj} values from individual tonal targets (e.g., $k=1$, $j=1$, or P11 in Fig. 1); $\varepsilon^{m2}_{\zeta_n}$, those by the mean 0.0089 in this experiment.

It is clear from Fig. 7 that (1) lexical tones in isolation are rather stable in a statistic sense; (2) there exist considerable internal differences among these tonal targets in F_0 in the left panel; and (3) the remains (i.e., extracted intonation components) after eliminating the citation values from these tonal targets become flat near $\zeta(t)=\zeta_0$ in the right panel.

Figure 8 displays the extracted intonation components from the 504 trisyllable words using the separation algorithm, where time intervals between two targets are normalized as equal intervals. It should be noted that: (1) The Tone 3 sandhi is treated as Tone 2, and (2) there are two regular tonal targets for each tone except for Tone 3. Tone 3 has two targets P31 and V31 (see Fig. 1) at the first syllable position, V31 and P32 at the last, but only one target (V31) at the middle. (3) Each tonal target takes its citation value $f_{c_{kj}}$ in

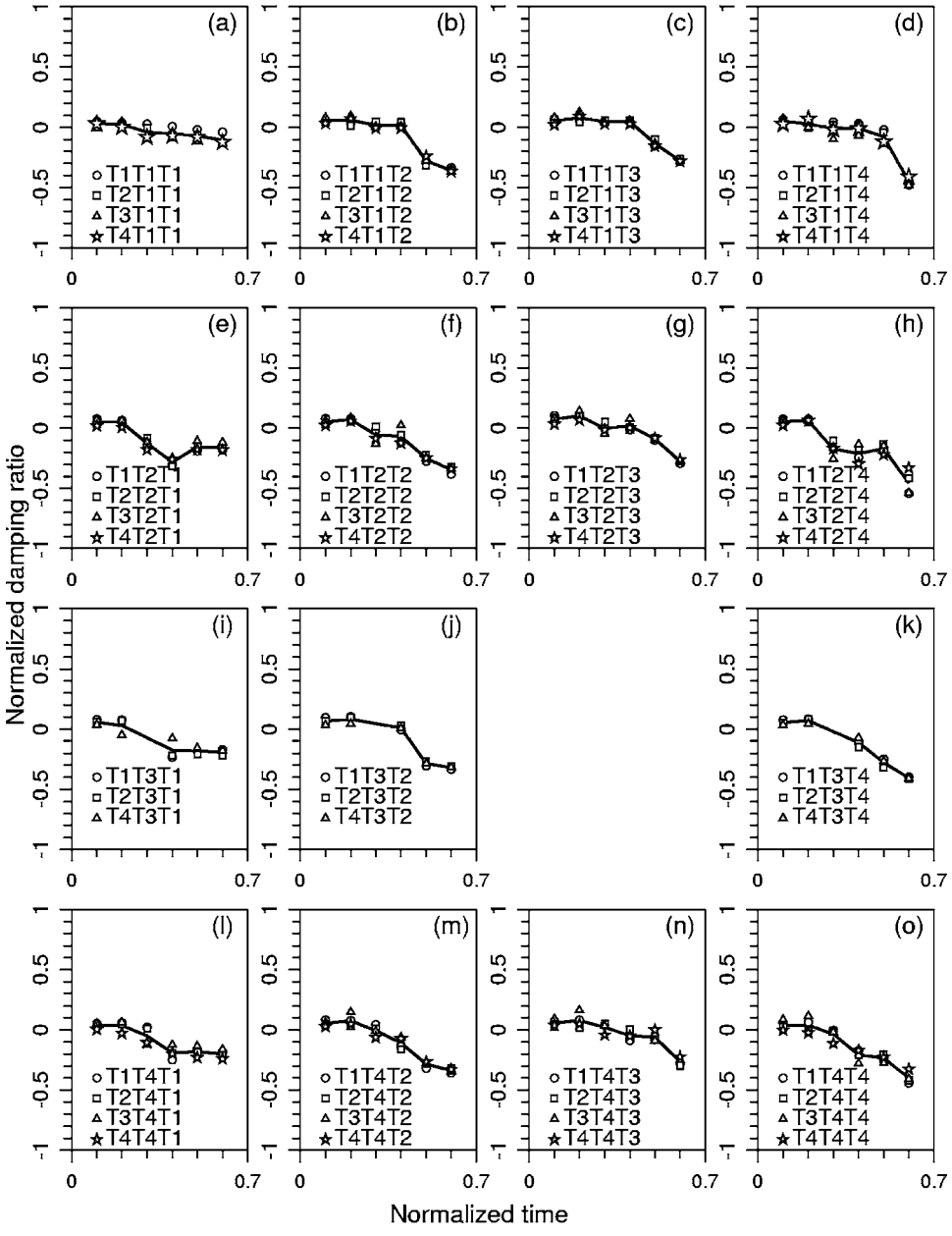


FIG. 8. Intonation components extracted from 504 three-syllable words covering all kinds of lexical tone combinations. Each symbol (circle, square, triangle, or star) indicates the means of several samples with individual tone combinations. The continuous lines in a panel indicate the mean of all the samples with the tone combinations superimposed on the panel. T1, T2, T3, and T4 indicate Tones 1 to 4, respectively.

Table I, $k=1,2,3,4$; $j \in \{1,2,3\}$, except for $f_{c_{42}}$ (V41 in Fig. 1). If Tone 4 is followed by another Tone 4, $f_{c_{42}}$ is replaced by $f^{h}_{c_{42}}$. Otherwise, it takes $f^{*}_{c_{42}}$. That is, $f_{c_{42}}$ itself is not used in the experiment. In each panel from (a) to (o), only the first tone changes from T1 to T4, where T*k* indicates Tone *k*; each symbol (circle, square, triangle, or star) indicates the individual mean of normalized ζ_{kj} values of the samples that have an identical tone combination [e.g., T1T1T1 in (a)]. The continuous lines in a panel indicate the mean of these normalized ζ_{kj} values from all the tone combinations shown in the panel [e.g., T1T1T1, T2T1T1, T3T1T1, and T4T1T1 in (a)]. The RMSEs between observed and resynthesized F_0 contours are listed at row "Words" of Table V. In this case, $\varepsilon^{m1}_{\zeta_n}$ corresponds to the RMSEs relative to the individual means indicated by individual symbols, such as circles, and $\varepsilon^{m2}_{\zeta_n}$ those to the lines.

It is clear from Fig. 8 that: (1) There exists a slight declination; it is universal (Ladd, 1983). (2) When focusing on the two tonal targets observed from the first syllables, all of the ζ_{kj} values, $k=1,2,3,4$ and $j \in \{1,2\}$, are almost consistent in a statistical sense and are comparable with those of isolated syllables. This result provides a strong experimental basis for supporting the proposed method in that the effect of lexical tones on observed F_0 contours can be eliminated with reference to their citation forms. (3) The final downdrift is tone dependent: A strong effect on Tone 4, a weaker effect on Tones 2 and 3, and almost no effect on Tone 1. (4) The contextual tonal effects are systematic in the results, which agree with the carryover and anticipatory effects described in Xu (1997) or the downstep phenomenon found in nontonal languages; Tones 2, 3, and 4 have a low target. In addition, there exists an "extra" effect related to Tone 2 as clearly shown in Fig. 8(e). It is not yet clear whether it is regular or speaker dependent. The results also indicate that the effects from individual factors on the extracted intonation components in ζ_n are additive. Take it as an example to compare the respective means of these ζ_n values from Tone 3 [the third triangle in Figs. 8(i)–8(k), corresponding to V31 in Fig. 1] among the following contexts: T4T3T1, T4T3T2, and T4T3T4. Mean ζ_n is quite similar between T4T3T1 in (i) and T4T3T4 in (k), but it is a little higher in T4T3T2 in (j) due to adding the slight anticipatory effect of Tone 2 (Xu, 1997) to the mean ζ_n without the anticipatory effect.

2. Statements and questions

Four groups of limited "designed" sentences are used in this experiment, basically adopted from Shen (1990) studying Mandarin prosody. Each group includes two base sentences and a combination of them, thus resulting in 24 sentences. Each sentence was uttered twice by FL as statements and questions with the question particle *ma0*, thus there existed 48 utterances.

Figure 9 shows the extracted intonation components from the 48 utterances whose lexical tone structures are also superimposed on this figure. The circle sequences indicate results for statements, while the square sequences do so for questions; a symbol (circle or square) denotes the respective means from two samples. The analysis conditions are the same as the setup for analyzing the isolated words. The RMSEs between observed and resynthesized F_0 contours are listed at row "Sentences" of Table V.

It can be seen from Fig. 9 that: (1) Intonation components of questions are of a higher level than those of statements, and considerable differences mainly exist near the end of utterances, as exemplified in Shen (1990). (2) There exist local humps along with a smooth curve; this is clearly shown in Figs. 9(a)–9(c). In this aspect, the extracted intonation components can be compared to a "channel" with a locally "gradual enlargement" for a focus, as described in Kratochvil (1998) about Chinese intonation. (3) The primary targets (black squares) of the last Tone 3 in the questions shown in Figs. 9(g)–9(i) could reach those in corresponding statements; this phenomenon also appears at the nonprimary targets of the last Tone 2 in the questions shown in Figs. 9(e) and 9(f), but does not appear in Tones 1 and 4. This may indicate that Mandarin has a tone-dependent tune, too, like Tune 4 described in Thai language (Luksaneeyanawin, 1998). In spite of the contextual tonal effects, it still is seen that there exists a certain similarity among these extracted intonation components when focusing on either statements or questions.

3. Read greetings

Four conventional greetings below in read speech are analyzed to further demonstrate that the extracted intonation components may be of certain patterns in similar pragmatic environments.

1. *ni3hao3*. (Hello.)
2. *zen3me0yang4a0?* (How are you doing?)
3. *ni3 mang2ma0?* (Are you busy?)
4. *ni3 shen1ti3 hao3ma0?* (How are you feeling?)

All of these speech samples were produced by a female professional (WL). The vocal range is fixed at [100 Hz, 500 Hz], and the citation values of tonal targets by WL are presented in Table I.

The experimental procedure is roughly described as follows, taking the first sample as an example.

(i) Manually determine observed targets $(\hat{t}_i, \hat{f}_{0_i})$, for example, (0.07 s, 243 Hz), (0.19 s, 367 Hz), (0.36 s, 197 Hz), (0.45 s, 152 Hz), (0.64 s, 204 Hz) for sample *ni3hao3*.
(ii) Prepare the citation values for Tones 2 and 3 where Tone 2 is the surface tone of syllable *ni3*: V21 (250 Hz), P22 (355 Hz), P31 (235 Hz), V31 (200 Hz), P32 (280 Hz) taken from Table I.
(iii) Compute ζ_{n_i} using the separation algorithm (Sec. III C 3) fed with the input data.

Figure 10 shows the observed F_0 contours ("+" sequences where the short vertical lines across them indicate the target positions) in the upper panels from (a) to (d) and corresponding extracted intonation components (lines through the circles) in the lower panels from (e) to (h). By comparing the upper panels with the lower panels, while the F_0 contours largely undulate due to underlying syllabic

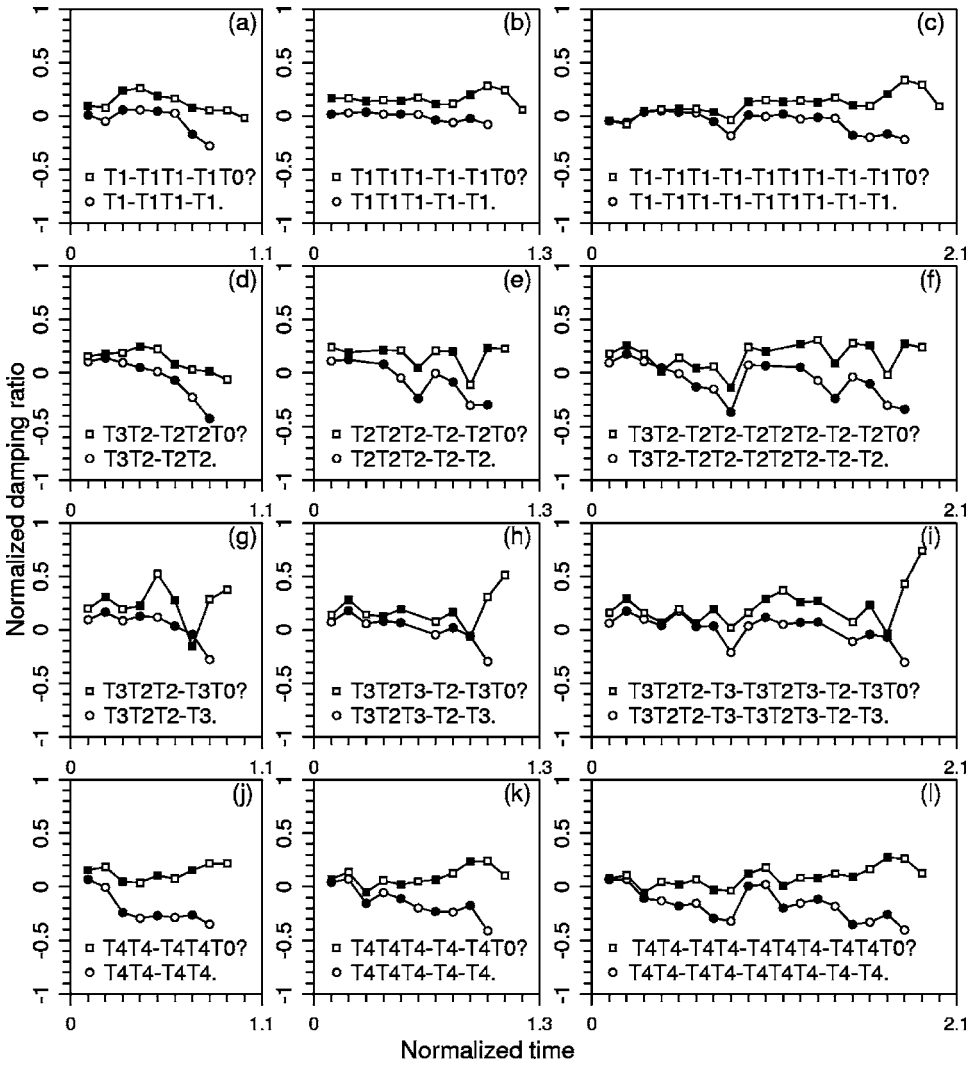


FIG. 9. Intonation components extracted from 24 statements and 24 questions; the circle sequences (means of two samples) indicate those extracted from the statements; and the square sequences do so for the questions.

tones, the intonation components involved in these examples are probably identical in that they have an identical pattern: "rising-[level]-falling level."

The main characteristics among the four samples are the last "falling-levels." The "falling" indicates the nucleus of sentence intonation, while the "level" indicates that the speaker is producing the complete forms of the last lexical tones in each greeting, probably to show her politeness or due to read speech. This intention is partly observable from Tone 0 in *ma0* (a question particle) in Fig. 10(c); Tone 0 transfers part of the preceding tone's patterns. Tone 0 continues the tendency of F_0 movements of the preceding Tone 2 to a target (corresponding to P22 in Fig. 1) and then approximately maintains it until near to the end. Compared with the findings on intonation phenomena in other languages, the extracted intonation components are similar to the high-low pitch accents used for descriptions of Japanese intonation (Shirai and Abe, 2000) or to the so-called "hat pattern" that is usually used to describe intonation patterns in some European languages, as exemplified in 'tHart *et al.* (1990).

These results indicate that through separating intonation components from observed F_0 contours by the tone transformation technique, it is promising to discover the regularity of intonation patterns, at least in the particular cases. The experimental results partly convince us that Mandarin intonation may have similar contour and range characteristics as the findings in other languages that, for example, are abstractly described in Hirst and Cristo (1998). Further work is needed in this aspect.

4. Intonation conversion

The constrained tone transformation technique can also be applied to measure intonation variation between two observed F_0 contours with an identical lexical tone structure. One F_0 contour is used as a *norm* (a normative form of a lexical tone structure), and the other is regarded as a *variant* (an observation form of the lexical tone structure). The intonation variations from the norms to the variants can then be measured by the tone transformation technique, too. This consists of transforming target values in pairs from the norms to the variants. These targets are intended to sparsely specify the prosodic contributions to the F_0 contours and are aligned in time between norms and variants in terms of the elements of the lexical tone structures. Also, this experiment is intended to demonstrate that the extracted intonation components can be reproduced in a perceptual sense.

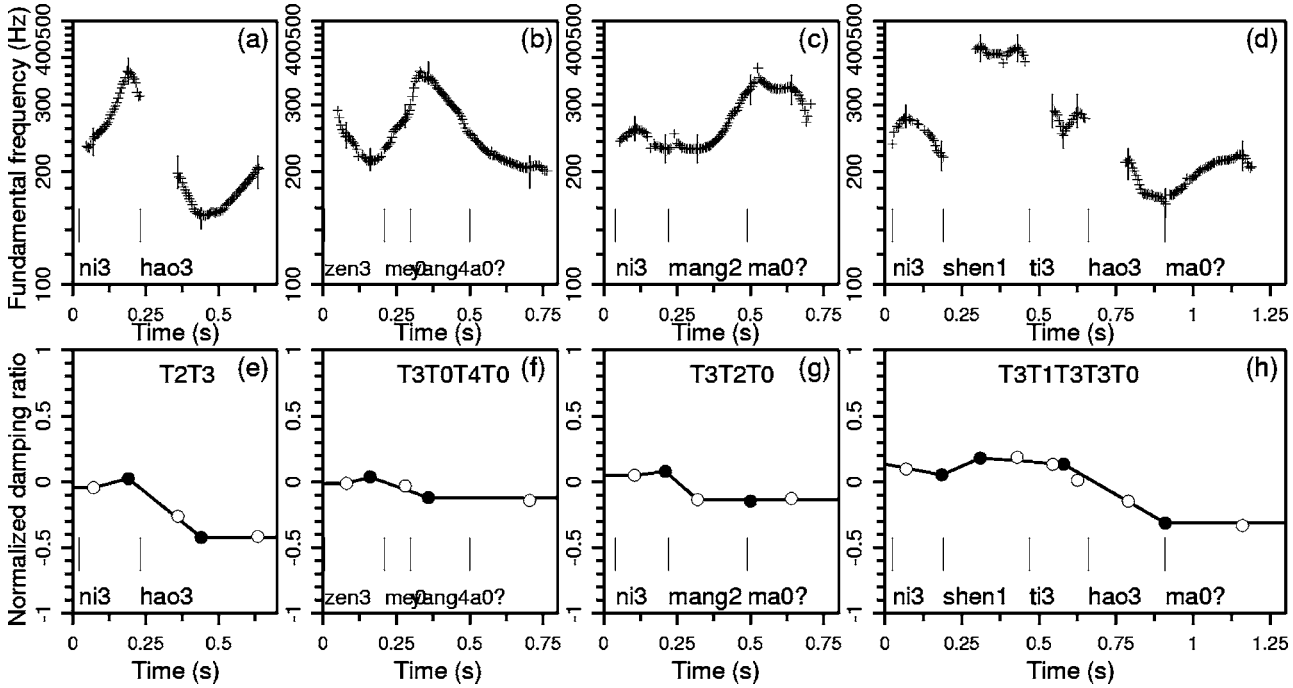


FIG. 10. F_0 contours ("+" sequences) for four conventional greetings: (a) *ni3hao3*. (Hello); (b) *zen3me0yang4 a0*? (How are you doing?); (c) *ni3 mang2ma0*? (Are you busy?); and (d) *ni3 shen1ti3 hao3ma0*? (How are you feeling?). The extracted intonation components are displayed in panels (e)–(h).

For this purpose, we select two pairs of speech samples below to perform intonation conversion between statements and questions.

(1) *guo4lu4ke4 zhao4 xiang4*. (A passerby takes photographs.)
(2) *guo4lu4ke4 zhao4 xiang4*? (Does a passerby take photographs?)
(3) *hong2bi2tou2 mei2 quan2*. (The Red Nose has no power.)
(4) *hong2bi2tou2 mei2 quan2*? (Has the Red Nose no power?)

Each pair, (1) and (2), or (3) and (4), is characterized by having identical lexical tone and syntactic structures, produced as a statement and a question by speaker FL. Her vocal range $[f_{0_b}, f_{0_t}]$ is fixed at [100 Hz, 500 Hz] as mentioned earlier.

Let $\hat{F}_{0_s}(t)$ and $\hat{F}_{0_q}(t)$ denote a pair of observed F_0 contours from statements and questions, respectively. The experimental procedure is roughly described as follows.

(i) Determine the tonal targets for sparsely specifying the observed F_0 contours.
(ii) Calculate intonation variation, for example, variation from $\hat{F}_{0_s}(t)$ to $\hat{F}_{0_q}(t)$ denoted by $\zeta_{s2q}(t)$, according to their sparser specifications; the approach is quite similar to the separation algorithm.
(iii) Superimpose $\zeta_{s2q}(t)$ (intonation variation) onto $\hat{F}_{0_s}(t)$ (statements), thus converting $\hat{F}_{0_s}(t)$ to $\tilde{F}_{0_q}(t)$ (synthetic F_0 contours) by combining Eqs. (3) and (4):

$$\tilde{F}_{0_q}(t) = T_{f_0}(T_\lambda(\hat{F}_{0_s}(t), \zeta_0), \zeta_{s2q}(t)). \tag{18}$$

(iv) Perform listening tests of re-synthesized samples with $\tilde{F}_{0_q}(t)$ and questions with $\hat{F}_{0_q}(t)$.

Obviously, either $\hat{F}_{0_s}(t)$ or $\hat{F}_{0_q}(t)$ can be used as a norm of the underlying lexical tone structures. The notion of variable norms has been suggested in Kratochvil (1998) for analyzing concrete intonation.

Figure 11 displays the observed ("+" sequences) and converted F_0 contours (triangle or circle sequences); the continuous lines indicate optimal approximations of the observed F_0 contours with $\zeta(t)$ fixed at ζ_0 for deriving their sparser specifications. The converted F_0 contour in a panel is generated by using Eq. (18) to superimpose respective intonation variations (the thin lines through the triangles and circles shown in Fig. 12) on corresponding norms ("+" sequences displayed in the panel).

Figure 12 plots intonation variations (triangle sequences) from the statements to the questions and those (circle sequences) from the questions to the statements. The thick lines show the main tendency of primary targets (black triangles and circles) of underlying syllabic tones. At a glance, the questions have a higher level and a rising end in comparison with the statements. The divergence of valleys (white triangles and circles) at the last syllables [i.e., *ke4* (guest) and *tou2* (head)] of the two subjects *guo4lu4ke4* (passerby) and *hong2bi2tou2* (the Red Nose) from the main tendency (the thick lines) may indicate the pivots, which are believed to mark the boundary between two prosodic phrases (Garding, 1987). The differences near the end between Figs. 12(a) and 12(b) may demonstrate the existence of tone-dependent tunes in Mandarin again.

By using the STRAIGHT-based analysis-by-synthesis technique (Kawahara *et al.*, 1999), synthetic contour $\tilde{F}_{0_q}(t)$ is

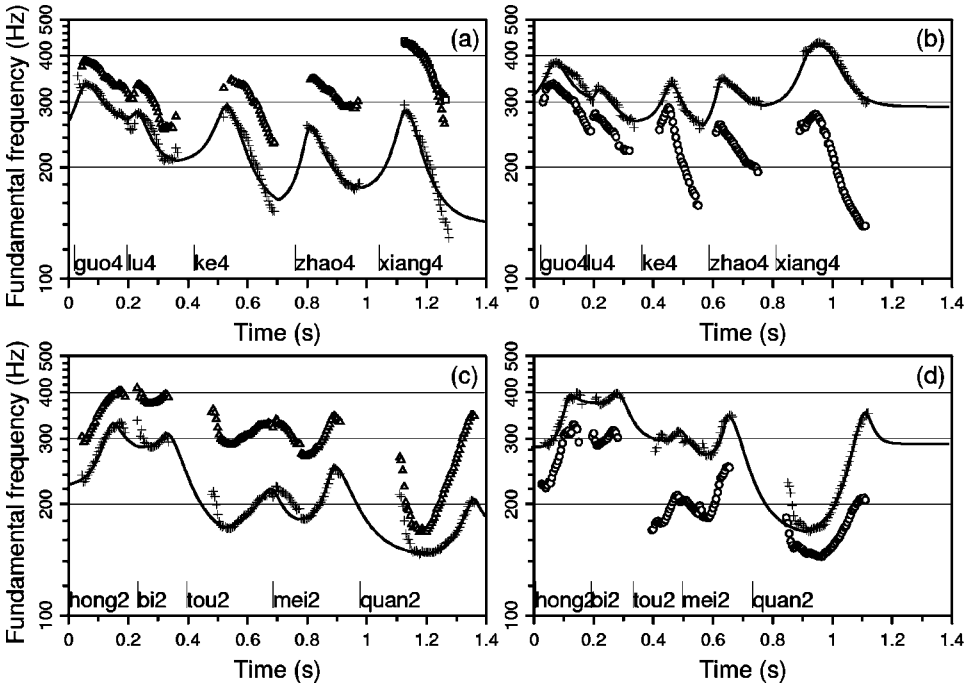


FIG. 11. Converted F_0 contours (triangle/circle sequences) by superimposing the intonation variations (the thin lines through triangle/circle sequences) shown in Fig. 12 on the corresponding observed F_0 contours ("+" sequences). The continuous lines are optimal approximations of the observed F_0 contours given by the model.

used to replace the natural values $\hat{F}_{0_s}(t)$ to resynthesize speech samples, such that the separated intonation variations are superimposed on the speech samples used as the norms; so does $\widetilde{F}_{0_s}(t)$ for $\hat{F}_{0_q}(t)$. An informal listening test was conducted on both the resynthesized and original samples for rating their prosody similarity. Basically, perceptual results show that the prosodic characteristics of the resynthesized samples are indistinguishable from the originals except for two points: (1) Due to no adaptation of temporal patterns for the conversion from statements to questions, the converted questions have a relatively slow speech rate. Consequently, they sound less natural than the questions with the observed F_0 contours $\hat{F}_{0_q}(t)$. (2) The resynthesized sample with the triangle sequences shown in Fig. 11(a) is perceptually more like a confirmation (meaning: The passerby does take photographs.) rather than a question (Does a passerby take photographs?). This is due to the divergence between the approximations given by the model and observed F_0 values at the end of the statement. There exists a considerable final downdrift in the statement [in Fig. 11(a)], while the estimated intonation variations, which are based on the sparser specifications of the observed F_0 contours (upon their approximations), did not cover the final downdrift. To confirm this point, we use the approximations given by the model [the continuous lines shown in Fig. 11(a)] to substitute the observed F_0 contours ["+" sequences in Fig. 11(a)] during the scope of the last syllable, and then obtain an alternative synthetic contour indicated by the square sequences in Fig. 11(a). As expected, the corresponding resynthesized sample clearly is a question.

C. Experiment 3: Justification of the two assumptions

1. Speech samples

The validity of using this kind of targets-based sparser specification for the observed F_0 contours can be tested by its ability to represent the F_0 contours involving rich information on tone and intonation. For this purpose, a double recording of a set of limited but well-designed sentences read by speaker FL is adopted from a Mandarin speech corpus (Wang *et al.*, 1996). In the sentence set, there are 12 base

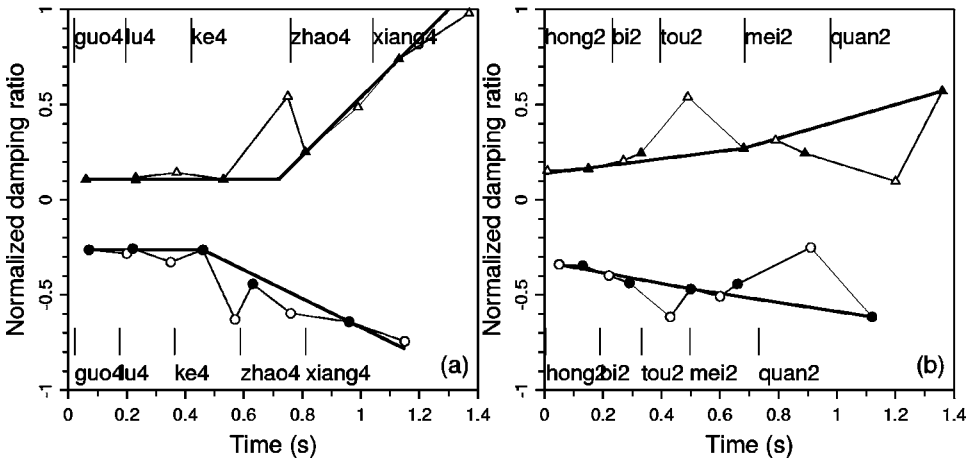


FIG. 12. Schematic diagram of the intonation variations between the questions and statements shown in Fig. 11. The thin lines through the triangle sequences indicate the estimated intonation variations from the statements to the questions, and the circle sequences do so from the questions to the statements. The thick lines indicate the main tendency of the primary targets (black triangles and circles) of underlying syllabic tones.

sentences subdivided into three types. Each type includes four sentences with the same number of syllables and identical grammatical structures. Type 1 comprises four syllables with a subject-verb (S-V) structure; Type 2 has five syllables with a subject-verb-object (S-V-O) structure; Type 3 combines Types 1 and 2, i.e., Type 2 is added to Type 1 as its sentential object, yielding nine syllables with an S-V-O structure. The 12 base sentences then constitute 72 sentences by themselves and them embedded in 11 typical question structures. The question structures include unmarked questions, yes-no questions, alternative questions, and all the wh-questions. These sentences were originally designed by Shen's (1990) study on Mandarin prosody.

2. Method

One hundred and forty-four F_0 contours are extracted from these utterances by Tempo [a component of STRAIGHT (Kawahara *et al.*, 1999)] at a frame length of 5 ms for interpolations, and the vocal range $[f_{0_b}, f_{0_t}]$ is fixed at [100 Hz, 500 Hz], a slight extension of the frequency range measured from the 144 F_0 contours.

The parameter extraction is described as follows. The observed F_0 contours are first automatically analyzed by the algorithm (Ni and Kawai, 2003) so as to estimate the underlying tonal F_0 peaks. Because of the limited performance of automatic peak detection (94% accuracy), these estimated tonal F_0 peaks are then manually adjusted by a visual inspection. Accordingly, 1448 peaks are determined for the speech sample set. With reference to these peaks, parameters Δt_{x_i} and $\Delta\lambda_{x_i}$, $x \in \{r, f\}$, for each utterance are then reestimated, thus turning out a set of model parameters together with these peaks, hereafter referred to as *baseline specifications*. Furthermore, there exist 804 valleys in the sample set, which are determined by first using the method described in Sec. III C 2 and then manually checking them with reference to underlying lexical tone structures for making these estimated tonal valleys aligned with the tonal F_0 valleys in time as accurate as possible. Assume that a manual check of these tonal targets is allowed with the specific experimental purpose: Tonal targets illustrated in Fig. 1 are sufficient for sparsely specifying the F_0 contours. A set of model parameters is then estimated from these peaks and valleys using Eq. (9), henceforth referred to as *sparser specifications*.

Both error analysis and listening tests are employed to rate the similarity between the resynthesized and natural F_0 values. In the first place, every resynthesized F_0 value in a vocal frame is compared with the corresponding natural F_0 values such that the absolute errors between them are calculated. The average absolute errors are then calculated for all the 144 F_0 contours.

3. Stimuli

Second, stimuli are designed for a perceptual experiment by use of the STRAIGHT-based analysis-synthesis technique (Kawahara *et al.*, 1999) to realize F_0 substitution. The F_0 values in each of the 144 utterances are replaced by (A) the F_0 contours resynthesized by the model using either the baseline or sparser specifications, and (B) the natural F_0 values. As a consequence, there exist 288 (=144×2) comparison pairs, one each from A and B.

TABLE VI. Results for prosody similarity between baseline and sparser specifications.

288(=144×2) comparison pairs	Score	Mean error (Hz)
F_0 values with sparser specifications vs natural F_0 values	1.89	6.38
F_0 values with baseline specifications vs natural F_0 values	1.93	5.93

4. Procedure

Three native subjects with normal hearing ability participated in this listening-test experiment. These stimuli were separately presented to each of the listeners through headphones in a silent room. The order in which A and B were presented to the listeners and the order in which comparison pairs were presented to the listeners were both randomized. After hearing a pair of stimuli, the listeners were asked to rate the similarity between A and B according to some factors, such as tones, intonation, and fluency at three levels: Very different, similar, and no difference. The listeners were allowed to hear the stimuli a few times before making a judgment.

5. Results and observation

Table VI outlines the experimental results. The average errors were 6.38 Hz in relation to the sparser specifications and 5.93 Hz to the baseline specifications. The column "Score" indicates the mean scores of rating the two sets of 144 comparison pairs times 3 (subjects) on prosody similarity: 0 (very different), 1 (similar), and 2 (no difference). In this experiment, no comparison pair was scored 0 by all the three subjects. This indicates that the resynthesized F_0 contours based on the sparser specifications are almost indistinguishable from the natural F_0 values. The baseline specifications produce slightly better performance than the sparser specifications do. This is because the baseline specifications produce closer approximations of these observed F_0 contours.

Table VII outlines the associations of observed targets, corresponding to the sparser specifications of the F_0 contours, with the underlying lexical tone structures, where Tone

TABLE VII. Distribution of tonal targets to lexical tones with the impact of sentence intonation.

Tone type	Tone count	Peak (1448)	Valley (804)	Rate (%)
Tone 0	54	48	16	n/a
Tone 1	214	418	50	87.2
Tone 2	364	502	348	91.7
Tone 3	142	190	140	100
Tone 4	290	290	250	98.6
Others	n/a	0	0	n/a

3 sandhi has been counted into Tone 2. The third and fourth columns, respectively, indicate distributions of the 1448 peaks and 804 valleys to each of the five tones, as well as others, if any. The fifth column indicates how many tones have full elements as illustrated in Fig. 1; the total tone numbers are listed in the second column. It is clear from Tables VI and VII that no extra target is necessary, given these targets required by the tone graphs, in the sense of retaining the essential characteristics of the prosodic contributions (of syllabic tones and sentence intonation). On an average, about 94% of the total number of Tones 1–4 are robust to the impact of sentence intonation.

Going into more detail about the 6% with tone exceptions, we found that most of the tones with distorted shapes occurring within those syllables that help constitute question structures, such as *zhao4mei2zhao4* (to take or not take). Such syllables include functional words *shi4* (to be), *hai2* (yet), *mei2* (not), *shen2* [part of the word *shen2me0* (what)] and others like *zhao4* (to take). Moreover, the tone distortion is mostly related to strong prominences due to the syntactic structure and/or speaker's attitudes. For example, in the question structure *zhao4mei2zhao4* (to take or not take), the culmination of a focus was found to coincide with the ending of the first syllable. Consequently, the first two lexical tones (i.e., Tones 4 and 2) were very heavily distorted, even changing direction. When the culmination of a focus coincided with a syllable carrying Tone 1, the contours of Tone 1 may appear to have valleys. As illustrated in Fig. 4(f) by the thin lines, the fourth Tone 1 has a valley. In this case, both peaks and valleys are counted into the experimental results. This is why valleys exist for Tone 1 in Table VII. From these points of view, most of the exceptions are related to the impact of sentence intonation. Therefore, the experimental results tend to justify the two assumptions for the proposed method.

6. Discussion

The proposed method deals with the acoustic correlates of tone and intonation involved in the F_0 contours, taking into account underlying lexical tone structures. The lexical tone structures are formed by concatenating a sequence of tone graphs. This consists of placing a series of targets along the time axis and these targets are always reached whereby truncated transition functions connect with minimum discontinuities. In this respect, this approach is close to the model of Pierrehumbert (1981), MOMEL (Hirst *et al.*, 2000), and others such as Garding (1987) by the use of a target-based representation of the F_0 continuum. The targets used here have some linguistic meaning because they directly specify the tone graphs to distinguish word meaning.

The proposed method involves a second level of control that consists of a simple mechanism to perform tone transformations. The tone transformation technique modulates the heights and ranges of these tone graphs (prototypical patterns) in F_0 according to the sentence modality and pragmatic constraints. This method is thus compared to the Tilt model proposed by Taylor (2000), the soft templates in Kochanski and Shih (2003), and others by modulating prototypical patterns in terms of the sentence modality and pragmatic constraints. A comparative benefit may be that this method is based on a physiologically driven model of an F_0 generation process and nicely formulated with the incorporation of the linguistic theory about Mandarin tone and intonation.

While an F_0 contour is treated as a coupling of both tone and intonation components, the coupling is not based on conventional suprapositon assumptions, such as in the Fujisaki model. Compared with the Fujisaki model, this method makes the intonation components hidden into the modulated tones, thus facilitating more latent transformations and avoiding the difficulty arising in decomposing F_0 contours into both tone and phrase components (Hirose *et al.*, 1994).

This tone transformation technique is now evaluated by using read speech samples and minimizing the effects from factors other than tone and intonation. However, it is inevitable to face this problem when such contextual factors as segmental and speaker's effects really exist. Also, it remains to be seen how the turning points are determined if the intonation components are piecewise linear. An initial thought to tackle these problems is to employ powerful statistic models and machine learning technology. By the use of appropriate statistic models with the expectation-maximization training, for example, it is possible to separate individual effects from linguistic and paralinguistic factors as well as such other affecting factors as segments and speakers on the remains, where the effects of lexical tones on the F_0 contours have been suppressed. An example of statistics-based models is exemplified in Chen *et al.* (2005). It looks promising in that the remains are represented by a series of very local targets in ζ_n, and more importantly, the effects of individual factors on ζ_n are of an additive property. Further work is needed in this aspect.

V. CONCLUSION

This paper presented a novel approach toward quantitatively separating and combining syllabic tones and sentence intonation in Mandarin, focusing particularly on the observed F_0 contours of read speech, undertaken in the spirit of analysis by synthesis. The problem is difficult because it is multidisciplinary, deeply involving linguistics, physiology, and acoustics. It remains to be seen how far the relation between tone and intonation observed from more speakers and spontaneous speech can be accommodated within the proposed approach. It also appears to us quite probable that the tone-transformation-based extraction of intonation components from observed tonal targets is sometimes sensitive to how these tonal targets are reliably estimated from acoustic data. However, a tool for quantitative analysis and synthesis of F_0 contours, such as that which we have sketched above, specifying explicitly different levels of parametric representation of tone and intonation, shall be interesting in the context of communicative speech synthesis and prosody analysis.

ACKNOWLEDGMENTS

This research was supported in part by the National Institute of Information and Communications Technology. We would like to thank the anonymous reviewers for their helpful comments and suggestions.

Abe, I. (**1980**). "How vocal pitch works," in *The Melody of Language*, edited by L. R. Waugh and C. H. Schooneveld (University Park Press, Baltimore, MD), pp. 1–24.

Chao, Y. R. (**1968**). *A Grammar of Spoken Chinese* (University of California Press, Berkeley, CA), 1–56.

Chen, S. H., Lai, W. H., and Wang, Y. R. (**2005**). "A statistics-based pitch contour model for Mandarin speech," J. Acoust. Soc. Am. **117**(2), 908–925.

Dusterhoff, K. E., Black, A. W., and Taylor, P. (**1999**). "Using decision trees within the Tilt intonation model to predict F_0 contours," in *the Sixth European Conference on Speech Communication and Technology*, pp. 1627–1630.

Fry, D. B. (**1979**). *The Physics of Speech* (Cambridge University Press, Cambridge, UK).

Fujisaki, H. (**1983**). "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The Production of Speech*, edited by P. F. MacNeilage (Springer, New York) 39–55.

Fujisaki, H. (**2004**). "Information, prosody, and modeling — With emphasis on tonal features of speech," in *Int. Conf. Speech Prosody 2004*, pp. 1–10.

Fujisaki, H., and Hirose, K. (**1984**). "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoust. Soc. Jpn. **5**, 233–242.

Garding, E. (**1987**). "Speech act and tonal pattern in standard Chinese: Constancy and variation," Phonetica **44**, 12–29.

Hirano, M. (**1976**). "Morphological structure of the vocal fold as a vibrator and its variations," Folia Phoniatr. **26**, 89–94.

Hirose, K., Lei, H. T., and Fujisaki, H. (**1994**). "Analysis and formulation of prosodic features of speech in standard Chinese based on a model of generating fundamental frequency contours" (in Japanese), J. Acoust. Soc. Jpn. **50**, 177–187.

Hirst, D., and Cristo, A. D. (**1998**). "A survey of intonation systems," in *Intonation Systems, A Survey of Twenty Languages*, edited by D. Hirst and A. D. Cristo (Cambridge University Press, Cambridge, UK), pp. 1–44.

Hirst, D., Cristo, A. D., and Espesser, R. (**2000**). "Levels of representation and levels of analysis for the description of intonation systems," in *Prosody: Theory and Experiment*, edited by M. Horne (Kluwer, Dordrecht), pp. 51–87.

Kawahara, H., Ikuyo, M. K., and Cheneigne, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**, 187–207.

Kawai, H., Toda, T., Ni, J., Tsuzaki, M., and Tokuda, K. (**2004**). "XIMERA: A new TTS from ATR based on corpus-based technologies," in *The 5th ISCA Speech Synthesis Workshop*, pp. 179–184.

Kochanski, G., and Shih, C. (**2003**). "Prosody modeling with soft templates," Speech Commun. **39**, 311–352.

Kochanski, G., Shih, C., and Jing, H. (**2003**). "Quantitative measurement of prosodic strength in Mandarin," Speech Commun. **41**, 625–645.

Kratochvil, P. (**1998**). "Intonation in Beijing Chinese," in *Intonation Systems, A Survey of Twenty Languages*, edited by D. Hirst and A. D. Cristo (Cambridge University Press, Cambridge, UK), 417–431.

Ladd, D. R. (**1983**). "Peak features and overall slope," In *Prosody: Models and Measurements*, edited by A. Cutler and D. R. Ladd (Springer, Berlin), 39–52.

Lee, L., Tseng, C., and Hsieh, C. (**1993**). "Improved tone concatenation rules in a formant-based Chinese text-to-speech system," IEEE Trans. Speech Audio Process. **1**, 287–294.

Lin, M. C. (**1995**). "A perceptual study on the domain of tones in standard Chinese," Chin. J. Acoust. **14**, 350–357.

Luksaneeyanawin, S. (**1998**). "Intonation in Thai," in *Intonation Systems, A Survey of Twenty Languages*, edited by D. Hirst and A. D. Cristo (Cambridge University Press, Cambridge, UK), 376–394.

Ni, J., and Hirose, K. (**2000**). "Experimental evaluation of a functional model of fundamental frequency contours of standard Chinese sentences," in *The 2000 Int. Symposium on Chinese Spoken Language Processing*, pp. 319–322.

Ni, J., and Kawai, H. (**2003**). "Tone feature extraction through parametric modeling and analysis-by-synthesis-based pattern matching," in *The 2003 Int. Conf. Acoustic Speech Signal Processing*, pp. 72–75.

O'Shaughnessy, D. (**1987**). *Speech Communication Human and Machine* (Addison-Wesley, Reading, MA), 380–412.

Pierrehumbert, J. B. (**1981**). "Synthesizing intonation," J. Acoust. Soc. Am. **70**, 985–995.

Ross, K. N., and Ostendorf, M. (**1999**). "A dynamical system model for generating fundamental frequency for speech synthesis," IEEE Trans. Speech Audio Process. **7**, 295–309.

Shen, J. (**1985**). "The range of Mandarin tones and intonation" (in Chinese), in *Experimental Pekinese Phonetics*, edited by T. Lin and L. J. Wang (Beijing University Press) 27–52.

Shen, X. S. (**1990**). *The Prosody of Mandarin Chinese* (University of California Publications).

Shirai, K., and Abe, M. (**2000**). *Recent Progress in Japanese Speech Synthesis* (Gordon and Breach, New York), 5–38.

Silverman, K., Beckman, M., Pitrelli, J., Ostendrof, M., Wighman, C., Price, P., Pierrehumbert, J., and Hirechberg, J. (**1992**). "ToBI: A standard for labeling English prosody," in *The 1992 Int. Conf. on Spoken Language Processing*, pp. 867–870.

Taylor, P. A. (**2000**). "Analysis and synthesis of intonation using the Tilt model," J. Acoust. Soc. Am. **107**, 1697–1714.

'tHart, J., Collier, R., and Cohen, A. (**1990**). *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody* (Cambridge University Press, Cambridge, UK).

Thorsen, N. G. (**1980**). "A study of the perception of sentence intonation—Evidence from Danish," J. Acoust. Soc. Am. **67**, 1014–1030.

Titze, I. R. (**1997**). "Regulation of fundamental frequency with a physiologically based model of the larynx," in *Speech Production and Language, In Honor of Osamu Fujimura*, edited by S. Kiritani, H. Hirose, and H. Fujisaki (Mouton de Gruyter, Berlin) 33–51.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (**2000**). "Speech parameter generation algorithms for HMM-based speech synthesis," in *The 2000 Int. Conf. Acoustic Speech Signal Processing*, pp. 1315–1318.

van Santen, J. P. H., Shih, C., and Möbius, B. (**1998**). "Intonation," in *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, edited by R. Sproat (Kluwer Academic, Dordrecht), 141–190.

Wang, R. H., Xia, D. Y., and Ni, J. (**1996**). "USTC95—A Chinese database," in *The 1996 Int. Conf. on Spoken Language Processing*, 1894–1897.

Wu, Z. (**1996**). "A new method of intonation analysis for standard Chinese: Frequency transposition processing of phrasal contours in a sentence," in *Analysis, Perception and Processing of Spoken Language*, edited by G. Fant *et al.* (Elsevier Science B. V., Amsterdam, Netherlands), 255–268.

Xu, Y. (**1997**). "Contextual tonal variations in Mandarin," J. Phonetics **25**, 61–83.

Xu, Y. (**1999**). "Effects of tone and focus on the formation and alignment of F_0 contours," J. Phonetics **27**, 55–105.

Interferometric studies of a piano soundboard

Thomas R. Moore and Sarah A. Zietlow
Department of Physics, Rollins College, Winter Park, Florida 32789

(Received 15 June 2005; revised 11 December 2005; accepted 13 December 2005)

Electronic speckle pattern interferometry has been used to study the deflection shapes of a piano soundboard. A design for an interferometer that can image such an unstable object is introduced, and interferograms of a piano soundboard obtained using this interferometer are presented. Deflection shapes are analyzed and compared to a finite-element model, and it is shown that the force the strings exert on the soundboard is important in determining the mode shapes and resonant frequencies. Measurements of resonance frequencies and driving-point impedance made using the interferometer are also presented. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2164989]

PACS number(s): 43.75.Mn, 43.40.At, 43.20.Ks [NHF] Pages: 1783–1793

I. INTRODUCTION

A complete understanding of how the dynamics of the modern piano creates its unique sound is unlikely without a thorough understanding of the soundboard. Probably the most important parameter associated with the soundboard is the impedance at the point where the strings meet the bridge, and there are several reports of investigations of the dependence of the driving-point impedance on frequency.[1–6] However, there are parameters beyond the mere value of the impedance at the terminating point of the strings that are important. Of particular interest are the soundboard deflection shapes. These shapes have been reported to be similar to the shapes of the normal modes for frequencies below approximately 200 Hz; above this limit the resonances of the soundboard are believed to be broad and closely spaced so that the deflection shapes do not necessarily resemble the mode shapes.[7,8] Understanding the deflection shapes is important not only because they provide some insight into the physical basis for the impedance structure at the bridge, but also because they can provide insight into effects, such as acoustic short circuiting, which may not affect the driving point impedance but may significantly affect the sound perceived by the listener.

The most common methods for determining the deflection shapes of objects include observing Chladni patterns, measuring the acoustic power as a function of position, making impact measurements, and performing holographic or speckle pattern interferometry. However, each of these techniques is difficult to use in determining the deflection shapes of a piano soundboard.

In principle Chladni patterns are simple to create, but they are extremely difficult to obtain on a fully assembled piano for several reasons. Although some Chladni patterns have been obtained on a soundboard isolated from the piano,[3] the utility of this method is limited by the necessity to physically access the entire soundboard, the requirement that the soundboard be oriented horizontally, and the fact that there are ribs and other attachments that segment the surface.

Recently, acoustic measurements have yielded information on the shapes of the lowest modes of a piano soundboard; however, these measurements have very low spatial resolution.[9,10] Likewise, mechanical measurements of the deflection of a soundboard have been made by physically placing accelerometers on an isolated soundboard as it is struck. This technique is limited by the necessity for physical access to all points of the soundboard, and it is therefore not feasible to perform the measurements on a fully assembled piano. Furthermore, the measurement process can potentially change the resonance structure and the spatial resolution is quite low.[2,7,8,11]

Time-averaged holographic or speckle pattern interferometry both provide interferograms that can reveal the deflection shape of a vibrating object, and both techniques have been used often to determine deflection shapes of musical instruments. However, the size and mass of a fully assembled piano make both types of interferometry a very difficult undertaking. Specifically, these interferometric techniques require extensive vibration isolation. A good estimate of the requirement for vibration isolation is that the movement of the object due to ambient vibrations must be significantly less than one-tenth of the wavelength of the light used to image the object. This stability must be maintained over the entire time it takes to perform the experiment. The high center of mass, large surface area, and wooden construction of a piano make this level of isolation problematic. Even when enclosed within a soundproof room, and mounted on supports with active vibration control mechanisms, vibrations that are transmitted through the support structure usually make such large objects too unstable to be effectively studied using these techniques.

To investigate the deflection shapes of a piano soundboard, we have modified the common form of the speckle pattern interferometer so that ambient vibrations that are difficult to eliminate do not degrade the time-averaged interferogram. In fact, when using this interferometer the decorrelation of the speckle pattern actually increases the precision of the measurements. This modification, combined with a recently reported modification that reduces the complexity of the interferometer while simultaneously reducing the necessary laser power,[12,13] has enabled us to optically investigate the deflection shapes of a fully strung piano soundboard *in situ*.

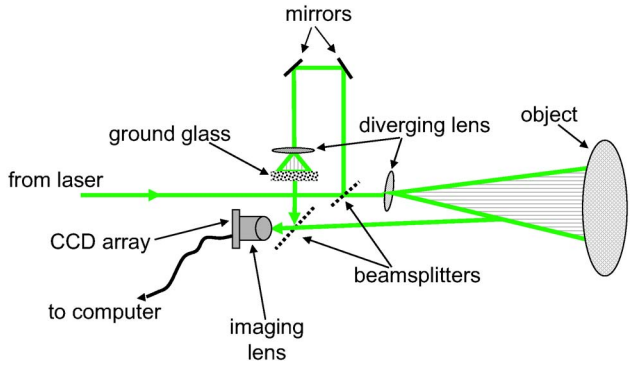


FIG. 1. (Color online) Diagram of the electronic speckle pattern interferometer.

In what follows we describe this interferometer, the theory of its operation, and its application to the study of a piano soundboard. We begin by discussing the most common form of the electronic speckle pattern interferometer and highlight the problems associated with using it to investigate the soundboard. We then describe a modification to the interferometer and present a theory showing that with this modification a slow decorrelation of the speckle pattern enhances the sensitivity. Next, we present some interferograms of the deflection shapes of the soundboard of a fully assembled piano that were made using this method, discuss their implications, and compare interferograms of the lowest modes with finite-element models. Finally, we demonstrate that data obtained using this interferometer can be used to determine the resonance curves of the soundboard, as well as make measurements of the driving-point impedance.

II. THEORY OF TIME-AVERAGED ELECTRONIC SPECKLE PATTERN INTERFEROMETRY WITH DECORRELATION

The arrangement of the electronic speckle pattern interferometer used in our experiments is shown in Fig. 1. It is a modified version of the more common arrangement and a detailed description can be found in Refs. 12 and 13. To create an interferogram, a beam of coherent light is reflected from an object and imaged through a beamsplitter onto a charge-coupled device (CCD) array. A reference beam is directed through a delay leg and illuminates a piece of ground glass, which is also in the view of the imaging lens through reflection from the beamsplitter. To study harmonically vibrating objects, the image present on the CCD array is digitally stored prior to the onset of harmonic motion of the object. Once the object is executing harmonic motion, the image on the array is then subtracted from the initial image in real time.

This situation can be described theoretically by assuming that the object under investigation is moving with simple harmonic motion, with an angular frequency of ω_0 and an amplitude of Δz. Assuming that the size of the speckle is approximately the size of a single pixel element, the intensity of the light on a pixel is given by

$$I = I_r + I_0 + 2\sqrt{I_r I_0}\cos\theta, \tag{1}$$

where I_r and I_0 are the intensities of the reference and object beam, respectively, and

$$\theta = \theta_0 + \xi\sin(\omega_0 t). \tag{2}$$

In this equation

$$\xi = \frac{2\pi\Delta z}{\lambda}(\cos\theta_i + \cos\theta_r), \tag{3}$$

and θ_0 is the initial difference in phases of the two beams at the pixel element. In Eq. (3), θ_i and θ_r represent the incident and reflected angles of the object beam measured from the normal to the surface of the object under study, and λ is the wavelength of the light.

For the case considered here the integration time of the detector is long compared to the period of the motion; therefore, it records only the time-averaged intensity of the interference. Assuming that the intensities of the two beams are constant in time, the recorded intensity is given by

$$\langle I\rangle = I_r + I_0 + 2\sqrt{I_r I_0}\langle\cos[\theta_0 + \xi\sin(\omega_0 t)]\rangle, \tag{4}$$

where the angled brackets indicate a time average. Performing the time average explicitly reduces Eq. (4) to[14]

$$\langle I\rangle = I_r + I_0 + 2\sqrt{I_r I_0}\,J_0(\xi)\cos\theta_0, \tag{5}$$

where J_0 represents the zero-order Bessel function of the first kind.

The most common method of viewing the amplitude of displacement is to record an initial image before the onset of vibration of the object and subtract it from an image recorded subsequent to the onset of harmonic motion; the absolute value of the difference is then displayed on a computer monitor. In this case the intensity recorded by the first frame is given by

$$I_1 = I_r + I_0 + 2\sqrt{I_r I_0}\cos\theta_0, \tag{6}$$

and the intensity recorded after the onset of vibration of the object is given by Eq. (5). When the two are subtracted, the intensity of the pixel displayed on the monitor after the nth frame is given by

$$I_n = \beta_n|1 - J_0(\xi)|, \tag{7}$$

where β_n is a positive constant with a value that depends upon the relative intensities of the image and reference beams, as well as the details of the display.

For the interferometer to provide the results indicated by Eq. (7), the object must be stable enough such that the speckle pattern does not decorrelate in the time between obtaining the initial and final image. Therefore, the interferometer and the object under study are typically isolated from ambient vibrations.

When the object under investigation is large or flimsy, adequate isolation from ambient vibrations can become difficult. Any displacement due to motion of the structure upon which the object rests increases linearly with height, and an object with a high center of mass, such as a piano, enhances the effects of any slight motion. Even when the interferometer and the piano sit on tables with active pneumatic vibra-

tion isolation, and the entire apparatus is enclosed within a room without circulating air, we have found that the speckle can decorrelate on a time scale significantly less than 1 second. This decorrelation may occur even more quickly if the interferometer and the object under study are independently supported, which may be required to image large objects such as pianos.

To include the decorrelation due to ambient vibration, an extra term must be added to Eq. (2). While the ambient vibrations may contain many frequencies, it is most difficult to isolate a large object from vibrations with long periods; therefore, we assume that the period of the movement responsible for speckle decorrelation is significantly greater than the period of the oscillation of interest. Under this condition,

$$\theta = \theta_0 + \frac{2\pi}{\lambda}\eta \sin(\sigma t) + \xi \sin(\omega_0 t), \tag{8}$$

where η is the amplitude of the oscillation and it is assumed that $\sigma \ll \omega_0$.

If the integration time of the detector is short compared to σ^{-1}, the small angle approximation may be applied and the movement can be approximated as a slow linear movement over the integration time of the detector. In this case Eq. (8) becomes

$$\theta = \theta_0 + \sigma t + \xi \sin(\omega_0 t). \tag{9}$$

Equation (9) can also be written as

$$\theta = \theta_0 + \varepsilon \omega_0 t + \xi \sin(\omega_0 t), \tag{10}$$

where $\epsilon \ll 1$. If the integration time of the detector is short compared to σ^{-1}, the intensity of the light incident on a pixel of the recording array is given by

$$\langle I_n \rangle = I_r + I_0 + 2\sqrt{I_r I_0}\langle \cos[\theta_0 + \varepsilon \omega_0 t + \xi \sin(\omega_0 t)] \rangle. \tag{11}$$

When integrated, the function within the angled brackets in the above equation is a form of the equation known as Anger's function, which is a generalization of the Bessel function of the first kind.[14] When ϵ is an integer, Anger's function reduces to a Bessel function of order ϵ.

Since we have assumed that $\epsilon \ll 1$, we can approximate the time-averaged term in Eq. (11) as Anger's function with $\epsilon = 0$. In this case Eq. (11) reduces to Eq. (5). However, since the small-angle approximation used to obtain Eq. (10) is not valid for times that are not short compared to σ^{-1}, θ_0 cannot be considered to be constant over long periods of time. That is, although we may assume that ϵ is approximately zero for the purpose of evaluating Eq. (11), any nonzero value of ϵ will result in a time-varying value of θ_0 that cannot be ignored over time periods that are not short compared to σ^{-1}.

If the time between the collection of images in electronic speckle pattern interferometry is long compared to σ^{-1}, and the integration time of the detector is short compared to this value, subtracting the pixel value recorded after the onset of vibration from the value of the same pixel recorded before the onset of vibration will generally provide a nonzero result. This nonzero result will occur even when the object has not been intentionally set into motion. Therefore, speckle pattern interferometry yields no usable information about the harmonic movement of the object unless the time between images is short compared to σ^{-1}.

However, when the initial image and the final image are both recorded while the object is oscillating at frequency ω_0, rather than the initial frame being recorded prior to the onset of vibration, Eq. (11) describes the intensity of both images. Under these circumstances the intensity shown on the computer monitor is proportional to the absolute value of the difference between the pixel values in the mth and nth images, i.e.,

$$I_{mn} = \beta_{mn}|J_0(\xi)|, \tag{12}$$

where in this case the value of β_{mn} depends not only upon the details of the monitor settings, but also on the value of the phase angle θ_0, which is slowly varying in time and therefore changes slightly between the mth and nth frames. That is,

$$\beta_{mn} \propto |\cos\theta_m - \cos\theta_n|, \tag{13}$$

where θ_m and θ_n represent the value of the initial phase angle for the mth and nth image, respectively.

If the object is too stable, such that $\cos\theta_m = \cos\theta_n$, the resulting interferogram will be uniformly black and yield no information despite any harmonic movement of the object. In this case it is necessary to subtract an image taken after the onset of vibration from one taken before vibration begins for the interferogram to provide any useful information. Likewise, if the movement that produces the decorrelation is not well approximated by a linear function over the integration time of the detector, then Eq. (12) is not valid.

In the case where there is a slow decorrelation of the light, such that over the integration time of the detector $\cos\theta_m \neq \cos\theta_n$ and $\epsilon \ll 1$, then the brightness of the pixel displayed on the screen will be proportional to $|J_0(\xi)|$, and lines of equal displacement will occur when

$$J_0(\xi) = 0, \tag{14}$$

where ξ is defined by Eq. (3). Averaging over several frames will eliminate the possibility of a spurious result due to $\cos\theta_m$ and $\cos\theta_n$ being equal simply by chance.

Note that Eq. (12) indicates that nodes in the soundboard will appear white on the monitor, while black lines indicate contours of equal amplitude of vibration. In the case where the initial frame is recorded before the onset of harmonic vibration, Eq. (7) indicates that areas of no movement appear as black, while white lines indicate contours of equal amplitude. Note also that the resolution of the system described by Eq. (12) is twice that described by Eq. (7), because Eq. (12) indicates that minima in the interferogram occur at every point where the Bessel function has a value of zero. Equation (7) indicates that minima occur only at the maxima of the Bessel function.

In the case where the object under investigation is stable so that $\cos\theta_m = \cos\theta_n$, a slow linear shift can be added to either the reference beam or the object beam; similarly, a slow physical motion of the object can be induced artificially. Either of these techniques will ensure that ϵ is nonzero but significantly less than unity, and will result in the increased

precision of the interferometer. Yet, precision is seldom the problem when attempting electronic speckle pattern interferometry in the laboratory.

As noted above, large objects such as pianos are normally not stable enough to image using electronic speckle pattern interferometry due to the presence of small ambient motions transmitted through the support structure. However, if the ambient motion is slow compared to the driving frequency ω_0, the motion will decorrelate the speckle within the constraints outlined above. In this case, speckle pattern interferometry can be used to study the motion of these objects. That is, as long as there is some slowly varying motion of the object, even if the motion is caused by ambient vibrations that normally precludes the use of interferometry, useful interferograms can be obtained.

Before concluding this section we note that assuming $\epsilon \sim 0$ ensures that we can approximate the time average in Eq. (11) as a zero-order Bessel function of the first kind. However, this is not a necessary approximation since Eq. (11) can be calculated explicitly for any value of ϵ. Interferograms can be obtained even for large values of ϵ; however, knowledge of its value is necessary to make measurements of the amplitude of the motion. We address this issue further in Sec. IV.

III. EXPERIMENTAL ARRANGEMENT

All of the experiments reported here were performed on a Hallet and Davis piano manufactured circa 1950. The piano was a spinet model, with a soundboard measuring approximately 1.4×0.63 meters. The soundboard was of varying thickness, as is usually the case; however, precise measurements of the thickness were not possible without significantly altering the piano and were therefore not made. The profile of the bottom of the soundboard could be unambiguously measured, and it was found that the thickness varied from 4.4 ± 0.1 to 6.9 ± 0.1 mm.

The soundboard of the piano was made of solid spruce, with the grain direction being 32° from the horizontal, diagonal from the upper left to lower right as viewed from the back of the piano. There were 12 ribs made of pine and glued to the back of the soundboard in an orientation perpendicular to the grain. The ribs had a width of approximately 25 mm, were spaced approximately 80 mm apart, and ranged in length from approximately 217 to 725 mm.

There were two bridges on the front of the soundboard over which the strings passed. A treble bridge, approximately 1.12 m in length was placed diagonally from the lower right to the upper left, as viewed from the back. A bass bridge, approximately 0.37 m long, was attached near the bottom of the soundboard. The edges of the soundboard were sandwiched between the case and retaining timbers, resulting in the edges being strongly clamped. Figure 2 contains a drawing of the soundboard, as viewed from the back, indicating the location of the bridges and orientation of the ribs.

Except where it is explicitly noted below, the piano soundboard remained completely strung and attached to the piano frame. No modification of the structure of the piano was made except that the hammers were removed for easier access to the treble bridge. The strings were all damped in two or three places by weaving strips of cloth between them to ensure that vibrations of the strings in no way affected the soundboard vibrations.

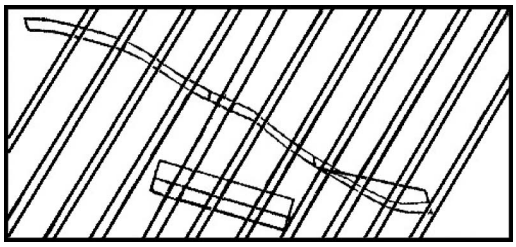


FIG. 2. Drawing of the orientation of the bridges and ribbing of the soundboard as viewed from the back. The bridges are attached to the front of the soundboard and the ribs are attached to the back.

The piano was mounted on an optical table with active pneumatic vibration-isolating legs, and attached to the table with a 3-in. nylon strap. The back of the piano was painted white to efficiently reflect light.

The interferometer was made from discrete optical components mounted on a separate actively isolated optical table. The entire experimental arrangement, with the exception of the laser, was contained within a $10 \times 12 \times 7$ ft. room, which was tiled with anechoic foam on all surfaces except for the portions of the floor that supported the optical tables.

The laser used to illuminate the piano was a frequency-doubled $Nd:YVO_4$ laser with a wavelength of 532 nm and a maximum power of 5 W. It was mounted outside of the anechoic room on an optical table with active pneumatic vibration isolation. The light entered the anechoic room through a small hole in the wall. A commercial CCD camera with a 768×494 pixel array and the standard 30-Hz frame rate was used to record the images. The illuminating beam and the imaging system were oriented perpendicular to the soundboard, so that $\cos \theta_i$ and $\cos \theta_r$ were both very close to unity.

Since the mass of the tables and the piano exceeded 1000 Kg, and the stability required for interferometry restricts the relative movement of the interferometer and piano to less than 50 nm, the floor of the chamber was not made of a suspended wire mesh as is common. Rather, the floor of the chamber was a portion of a concrete slab that comprised the floor of the building. Although the interferometer and the piano were mounted on tables with active vibration isolation, and both were housed in a chamber tiled with anechoic foam, there was a slight independent motion of the two tables at low frequencies. This motion was transmitted to the tables through the floor, with the peak transmissibility occurring at approximately 1 Hz.

Under normal circumstances the motion of the tables decorrelates the speckle and precludes the use of speckle pattern interferometry. However, since the method outlined above requires some decorrelation, the motion of the tables was advantageous. Thus, the technique described above allowed the use of speckle pattern interferometry in a situation in which it is normally impossible.

As outlined in the theory above, an image from the camera was digitally stored after the soundboard was set into

harmonic motion. Each subsequent frame was then digitally subtracted in real time from the previous image. Black lines appeared on the interferogram when the condition set forth in Eq. (14) was met.

During the experiments, vibrations of the soundboard were induced at an angular frequency of ω_0. The two motions, one harmonic with angular frequency ω_0 and one slowly varying so that it appeared linear over the integration time of the detector, were completely independent of one another. The slower vibrations occurred continually, regardless of the presence or absence of the driven harmonic motion.

Since the decorrelation time due to the ambient vibrations was on the order of the integration time of the CCD array, good interferograms were obtained by subtracting contiguous frames, thus allowing real-time viewing of the interferograms. Some higher frequency ambient vibrations resulted in noticeable noise in the interferograms, which was manifest as a series of lines unrelated to the harmonic motion of the piano. To eliminate these effects, and produce unambiguous interferograms for later analysis, up to 200 interferograms were averaged before the image was digitally stored. The total time of data collection for a single image, including postprocessing, was less than 10 s.

To obtain interferograms showing the deflection shapes of a piano soundboard, the apparatus described above was used to study the soundboard as it was driven harmonically. The driving force was provided by a speaker placed approximately 2 m from the soundboard. A high-quality function generator provided a sinusoidal signal, which was subsequently amplified and sent to the speaker. Typical deflections of the soundboard were less than a few wavelengths of the illuminating light (i.e., $\sim 0.1-2\ \mu$m) and required a sound intensity level on the order of 50 dB.

In some cases an electromagnetic shaker was used to drive the soundboard vibrations. The shaker was mounted on an adjustable magnetic base so that the driving mechanism could impinge upon any desired part of the soundboard.

IV. RESULTS AND ANALYSIS

A. Deflection shapes of a piano soundboard

Typical interferograms obtained using the apparatus described above are shown in Fig. 3. Three vertical braces and two carrying handles are visible in the interferograms, in addition to the soundboard and the rectangular frame to which it is mounted. The braces and handles are not physically touching the soundboard.

The interferograms in Fig. 3 show the deflection shapes of the soundboard at three different frequencies ranging from 219 Hz to 2.8 kHz. The total amplitude of the deflection at any point can be determined by counting the lines of equal displacement from a nodal point. Table I lists the displacement from the equilibrium position represented by each contour line.

From observations of the interferograms at frequencies between 60 Hz and 3 kHz, it appears that most of the resonances of the soundboard are broad and overlapping. However, the lower resonances appear to be significantly separated in frequency and therefore amenable to study. This conclusion has also been reached by others.[3,7,8] Interferograms of the deflection shapes of the lowest three resonances of the piano are shown in Figs. 4(a), 5(a), and 6(a). The resonant frequencies are shown in the second column of Table II. The bandwidth of the resonances are on the order of 20 Hz; however, the peak response is easily discerned by viewing the interferograms in real time while changing the driving frequency.

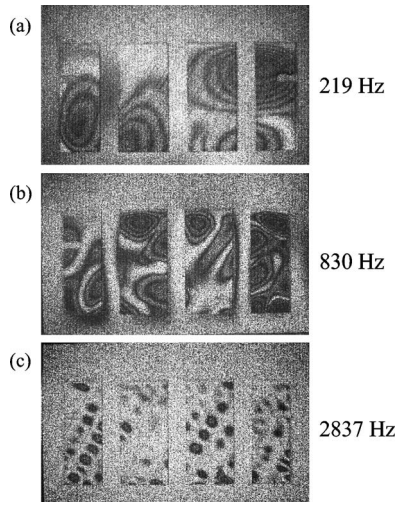


FIG. 3. Interferograms of the soundboard of a spinet piano. The vibrations are driven acoustically from a speaker placed approximately 2 m away.

The simplest model of a piano soundboard is that of an isotropic rectangular plate clamped at the edges. Leissa has reprinted the frequency parameters necessary for finding the resonant frequencies of rectangular plates with several aspect ratios.[15] The frequency parameter is defined as

$$\gamma = 2\pi f a^2 \sqrt{\frac{\rho}{D}}, \tag{15}$$

where f is the resonant frequency, a is the length of the shortest side of the plate, ρ is the area density, and D is the rigidity of the plate, which is defined as

TABLE I. The total amplitude of vibration represented by the number of dark lines traversed from a nodal point in an interferogram. Each line represents a point where Eq. (14) is valid.

Contour	Displacement (nm)
1	102
2	234
3	366
4	499
5	632
6	765
7	898
8	1031
9	1164
10	1297

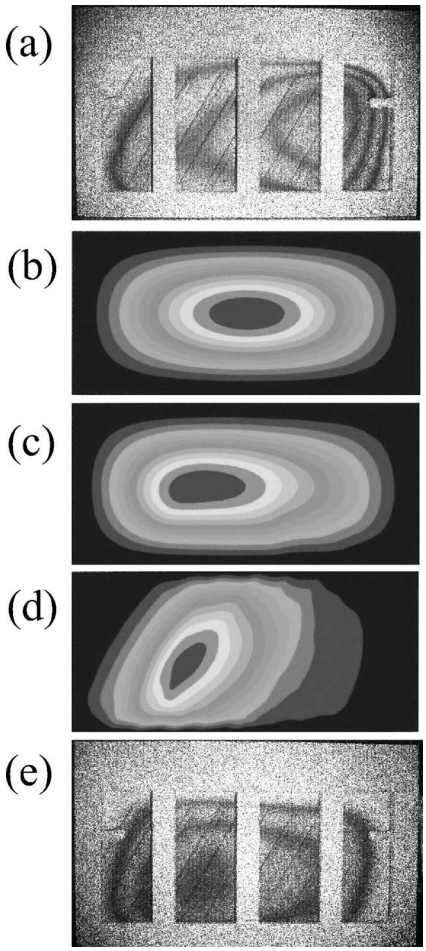


FIG. 4. Deflection shape of the first mode of the piano soundboard (a) observed interferometrically; (b) predicted by finite-element analysis of an isotropic soundboard; (c) predicted by finite-element analysis of an isotropic soundboard with bridges attached; (d) predicted by finite-element analysis of an orthotropic soundboard with ribs and bridges attached; and (e) observed interferometrically with no string tension. The resonant frequencies are listed in Table II.

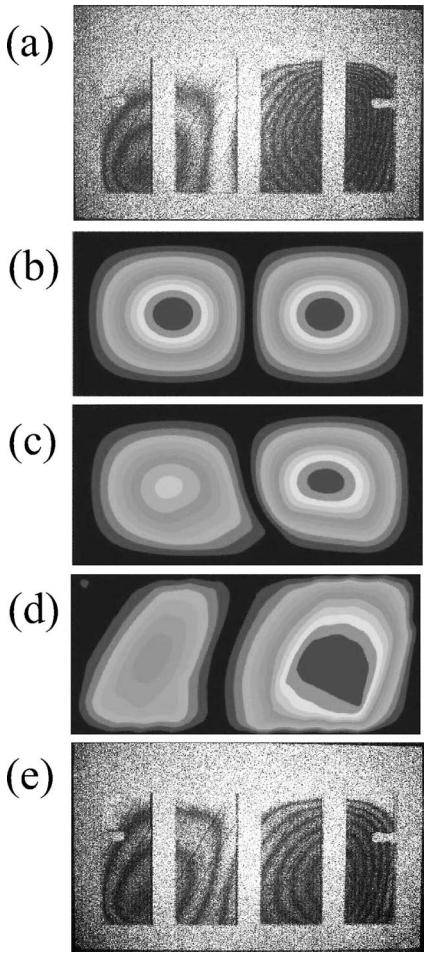


FIG. 5. Deflection shape of the second mode of the piano soundboard (a) observed interferometrically; (b) predicted by finite-element analysis of an isotropic soundboard; (c) predicted by finite-element analysis of an isotropic soundboard with bridges attached; (d) predicted by finite-element analysis of an orthotropic soundboard with ribs and bridges attached; and (e) observed interferometrically with no string tension. The resonant frequencies are listed in Table II.

$$D = \frac{Eh^3}{12(1-\nu^2)}, \tag{16}$$

where E is Young's modulus, h is the thickness of the plate, and ν is Poisson's ratio. Knowing the frequency parameter for any mode uniquely determines the frequency of that mode.

The aspect ratio of this particular soundboard is 0.45, which is very close to the aspect ratio of 0.5, for which the frequency parameters are reported by Leissa derived from work by Bolotin.[16] A good estimate of the correct value of the frequency parameters can be determined from these values by interpolation, which yields frequency parameters of 23.9, 29.9, and 40.7 for the first three modes of the soundboard.

Using Eq. (15), the frequencies of the lowest modes of an isotropic soundboard were calculated. The values for the density, Young's modulus, and Poisson's ratio were taken from the literature; all other parameters were measured.[17] These calculated frequencies are shown in the third column of Table II. This model is adequate to predict the approximate frequencies of the two lowest resonant modes of this soundboard; however, there are significant variances between the observed mode shapes and those of an isotropic rectangular plate clamped at the edges.

An analysis of the mode shapes shown in Figs. 4(a), 5(a), and 6(a) leads one to believe that the soundboard is not modeled well by a simple isotropic rectangular plate clamped at the edges, even though the resonant frequencies of the two lowest modes can be estimated using this model. Since the ribs are oriented in the direction perpendicular to the wood grain of the soundboard, and the spacing between them is much smaller than the wavelengths of the bending waves that are associated with the lowest modes in the wood, the soundboard should present an almost isotropic medium for the lower modes.[2] However, the soundboard under investigation exhibits some large-scale anisotropy, which can be deduced because the antinodes of all of the lowest three modes are noticeably shifted from their expected, symmetric positions.

Since the wavelength of the flexural waves in the soundboard at these low frequencies is on the order of the size of the soundboard, and there are no obvious small-scale aberra-

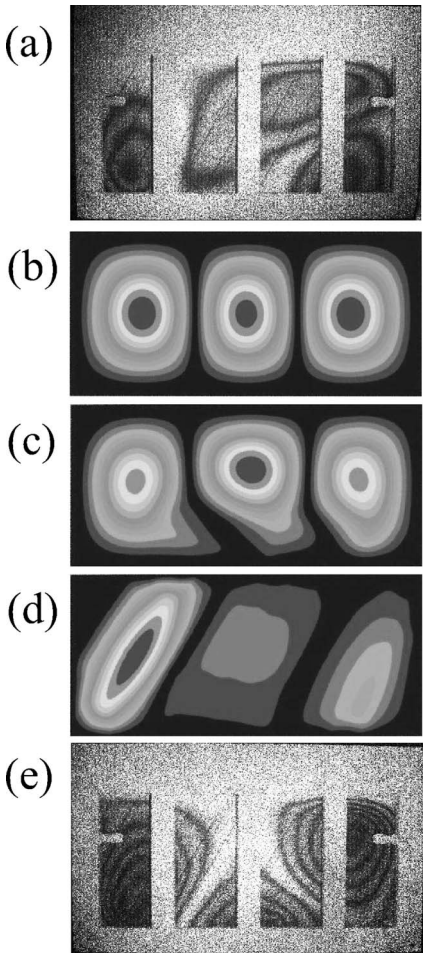


FIG. 6. Deflection shape of the third mode of the piano soundboard (a) observed interferometrically; (b) predicted by finite-element analysis of an isotropic soundboard; (c) predicted by finite-element analysis of an isotropic soundboard with bridges attached; (d) predicted by finite-element analysis of an orthotropic soundboard with ribs and bridges attached; and (e) observed interferometrically with no string tension. The resonant frequencies are listed in Table II.

tions in the mode shapes, it is likely that the cause of the asymmetry in the mode shapes is not due to small asymmetries in the soundboard. Indeed, it is likely that the observed asymmetries of the mode shapes are due to the presence of the bridges.

To test this hypothesis a finite-element model was constructed using a commercially available software program. The soundboard was assumed to be an isotropic board of sitka spruce with a thickness of 5.6 mm, which was the average thickness at the bottom of the actual soundboard. The model had approximately 39 000 elements and the edges were assumed to be clamped. The results of the model are shown in Figs. 4(b), 5(b), and 6(b). The resonant frequencies of the soundboard predicted by this model are found in the fourth column of Table II. As expected, the modes are symmetric and the predicted resonant frequencies are similar to those found analytically.

When the two bridges are added to the model the frequencies of the two lowest resonances change little, but the modal shapes of all three modes change significantly. The resonant frequencies of the soundboard predicted by this model are found in the fifth column of Table II. Figures 4(c), 5(c), and 6(c) show the shapes of the first three modes predicted by this model. Note that the antinodes and the nodal lines are shifted from the positions predicted by the simple model of an isotropic board. However, comparing these predictions with the interferograms shown in Figs. 4(a), 5(a), and 6(a) reveals that these mode shapes do not compare well with the experimental data. In particular, the antinode of the first mode and the nodal line of the second mode are both shifted to the incorrect side of the center of the soundboard. Additionally, the predicted deflection shape of the third mode remains significantly different than the experimentally observed shape.

Despite the poor agreement between the observed and predicted deflection shapes, the frequencies of the modes predicted by this model agree quite well with the observed frequencies. In particular, the addition of the bridges to the model has a significant effect on the frequency of the third mode. However, the large-scale anisotropy of the mode shapes, which does not appear to be accounted for by the presence of the bridges, indicates that it may be necessary to include the orthotropic nature of the wood and the presence of the ribs in the model.

The predictions of the mode shapes from a model that includes the orthotropy of the wood, the ribs, and the bridges are shown in Figs. 4(d), 5(d), and 6(d). The resonant frequencies of the soundboard predicted by this model are found in the sixth column of Table II. Note that the predictions of this model show better agreement with the experimentally observed mode shape of the second mode than the previous model, but the poor agreement between the predictions for the first and third modes remains. Additionally, the predicted frequencies for all modes are significantly changed, showing poor agreement with the experimentally derived values.

From this analysis one may conclude that by adding complexity to the model, and thus making it more similar to the actual piano, the predicted resonant frequencies are

TABLE II. The measured and calculated resonant frequencies (in Hz) of the soundboard. See the text for details.

Mode	Measured (±1 Hz)	Analytical	FEA isotropic	FEA isotropic with bridges	FEA orthotropic with ribs and bridges	Measured without strings (±1 Hz)
1	112	93	99	92	74	80
2	129	116	123	126	90	110
3	204	158	168	198	101	170

shifted further from the actual frequencies. Furthermore, the mode shapes predicted by the model are significantly different from those observed experimentally.

The poor agreement between the model and the interferograms leads one to believe that one or more important parameters are still missing from the model, despite the fact that all of the obvious physical parts have been included. We have determined that one of these parameters is the force exerted by the strings on the bridge.

As noted by Conklin, the total force of the strings on a modern piano is approximately 200 kN, and the bearing force on the bridges is between one-half and 3 percent of the string force.[3] Thus, the total force on the bridges is in excess of 2000 N. Although this force is distributed along the length of the bridges, it is unlikely that a force of this magnitude will not have an effect on the motion of the soundboard.

To determine if the bearing force of the strings does indeed have a large enough effect on the frequencies and mode shape to account for the differences between the model and the experimentally derived shapes and frequencies, the tension of each string was reduced until the string lay slack. Once the strings were no longer exerting a force on the soundboard, the experiments described above were again performed. The resulting interferograms are shown in Figs. 4(e), 5(e), and 6(e), and the measured resonant frequencies are shown in the final column of Table II.

Comparing Fig. 4(d) with Fig. 4(e), and Fig. 5(d) with Fig. 5(e) indicates that the finite-element computer model of the soundboard predicts the mode shapes and resonant frequencies quite well, as long as there is no significant pressure exerted by the strings. From this we may conclude that, although work on experiments and modeling that has been reported in the literature has largely ignored the pressure of the strings on the soundboard, it does have a significant effect on the mode shape and resonant frequencies of the lowest modes.

It is not clear, however, that the force due to the strings has a significant impact on the deflection shapes at higher frequencies. This can be seen by comparing the interferograms shown in Fig. 3, made with the strings at full tension, with those in Fig. 7, which were obtained with the soundboard vibrating at the same frequencies after the strings had been removed. The similarity of the interferograms indicates that the effects of the string force on the deflection shape is not as important at higher frequencies as it is at the lower frequencies.

Eliminating the pressure on the bridges due to the strings does have a significant effect on the shape and frequency of the third mode, but the mode shape predicted by the model still does not agree well with the interferogram of this mode. Currently, we do not know why there is such poor agreement between the model and the experiment for the frequency and shape of the third mode.

It is interesting to note that once the pressure of the strings is removed from the soundboard, the effects of the individual ribs becomes apparent at low frequencies. Note, for example, that the node of the second mode is parallel to the side of the soundboard when the bridges are under the pressure exerted by the strings, but without this pressure the node follows the direction of the ribs. This indicates that the soundboard of a piano can be modeled as an isotropic plate at low frequencies only when the pressure due to the strings is present. Evidently, it is not only the ribs that compensate for the orthotropy of the wood at low frequencies, but the pressure placed on the bridges by the strings appears to also help in this regard.

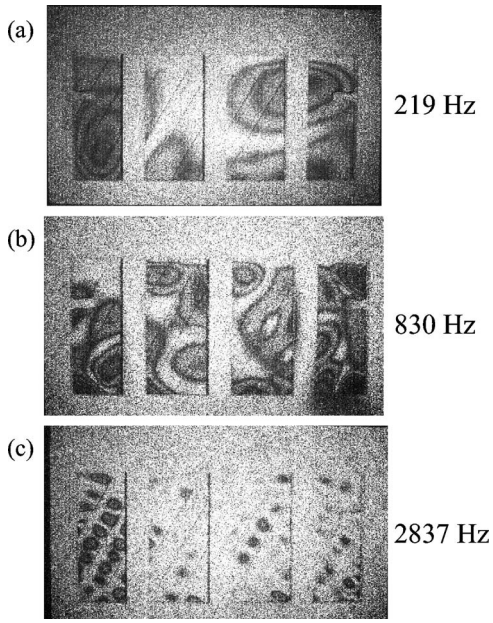


FIG. 7. Interferograms of the piano soundboard with the strings removed. The frequencies of vibration are the same as those in Fig. 3.

As noted above, at frequencies above those of the lower resonances, the pressure on the soundboard due to the strings appears to be less important. One reason for this may be that at higher frequencies the antinodes are small compared to the size of the soundboard, and except for the antinodes very near the bridges, the effect of the pressure on the bridge is unimportant.

In closing we note that, while the discussion above deals almost exclusively with the lowest two modes of the soundboard, the motion of the soundboard at these frequencies is extremely important in determining the overall tonal quality of the sound. For example, the asymmetry in the second mode, evident in Fig. 5(a), reduces the efficiency of acoustic short-circuiting between the two antinodes. This causes the lower frequencies to be radiated more efficiently than would be possible with a completely symmetric mode shape.

B. Comparison of the vibrational patterns of an acoustically driven and a point-driven soundboard

Although the deflection shapes of the soundboard are dependent only upon frequency when driven by a distributed acoustic signal, they can vary significantly when driven mechanically at a specific point. Since several resonances may be excited at any frequency, the position of the excitation will determine which resonances will be most efficiently excited. Therefore, excitation may occur at a node of a mode at one point but at an antinode at another, leading to the fact that the point of excitation will determine the deflection shape to a great extent.

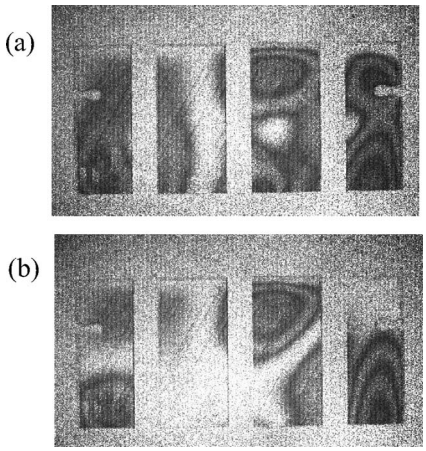


FIG. 8. Interferograms of the soundboard (a) when driven at a point on the bridge, and (b) when driven by a distributed acoustic source.

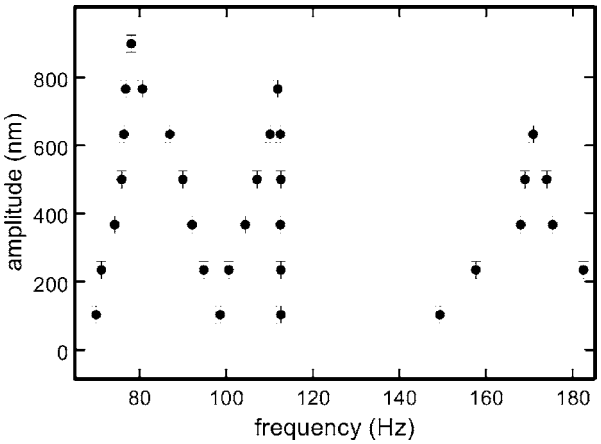


FIG. 9. Plot of amplitude of vibration vs frequency for the lowest three resonances of the piano soundboard.

To demonstrate the extent to which this can affect the deflection shape of the soundboard, an electromechanical shaker was placed so that it impacted the bridge near the terminating point of the C4 string on the treble bridge of the piano. Figure 8(a) shows the deflection shape of the soundboard when driven mechanically at a frequency of 328 Hz. For comparison Fig. 8(b) shows the deflection shape when the soundboard is driven acoustically at the same frequency by a speaker placed 2 m away. The difference between the two images in Fig. 8 leads one to ask if the shape of the bridge is actually designed to take advantage of the mode shapes of the soundboard.

Giordano has proposed that the shape of the bridge is designed to ensure that each string terminates near a location such that it cannot strongly excite a resonance in the soundboard at the fundamental frequency of the string.[4] This implies that each string terminates near a node in the soundboard, thus ensuring a slow transfer of energy to the soundboard from the string, and a commensurate lengthening of the decay time. We have investigated this conjecture by acoustically driving the motion of the soundboard at the fundamental frequency of each string using a speaker placed approximately 2 m away, and observing the deflection shape at the terminating point of the string associated with that frequency. In almost every case, the terminating point of the string lies extremely close to a node in the soundboard. Of all 88 strings, only the termination of the D5 string was not close to a node when the soundboard was driven at the fundamental frequency of the string. The termination of the D5 string was approximately midway between a node and antinode.

Preliminary experiments on other instruments indicate that not all pianos are designed so that the strings terminate near nodes in the soundboard at the fundamental frequency of vibration. Observations on a 6-ft. grand piano show that many of its strings terminate near antinodes.

C. Resonance measurements

Using the interferometer it is also possible to determine the resonance curve of the soundboard as long as a single antinode can be associated with an individual resonance. To do this it is only necessary to plot the amplitude of the deflection of one antinode of the mode as a function of frequency. Figure 9 contains such a plot of the lowest three modes of the soundboard without the strings.

The data shown in Fig. 9 were derived by driving the vibrations of the soundboard with a speaker placed approximately 2 m from the piano. The amplitude of the motion was determined using the interferometer as the driving frequency was varied. Using this technique it is possible to determine the frequency at which the modes begin to overlap significantly, thus setting an upper limit on modeling the soundboard as having individually isolated resonances. The data shown in Fig. 9 indicate that below approximately 200 Hz the modes are spaced widely enough to analyze them independently, although there is some overlap between the first and second modes far from resonance. Above this frequency the positions of the antinodes begin to shift noticeably with slight changes in frequency, indicating that there is significant overlap of the modes.

D. Measurement of the driving point impedance

The driving-point impedance determines much about the interaction of the string with the soundboard. It is defined as the driving force divided by the velocity, and to determine the driving-point impedance both of these quantities must be measured simultaneously. In the past, an impedance head that simultaneously measures the applied force and the acceleration has been used to determine the driving-point impedance; however, this arrangement has some disadvantages. The size and shape of these devices make it difficult to make measurements in the confined spaces of many pianos; therefore, only a limited number of measurements have been made on a fully assembled piano. Also, there have been reports of discrepancies in measurements that may be traceable to the construction and application of these devices.[1,6]

We have used an electronic speckle pattern interferometer to measure the driving-point impedance at several points on the soundboard, and at several different frequencies, without the use of an accelerometer. The interferometer was used to determine the velocity of the driving point while a shaker vibrated the soundboard at the point of interest. To measure

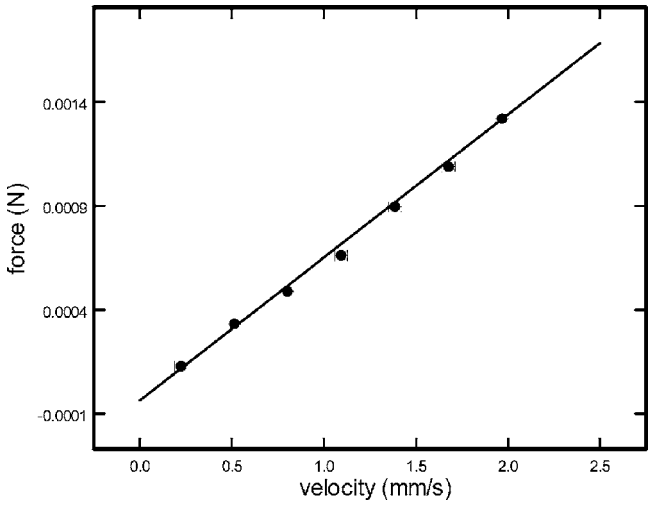


FIG. 10. Plot of the force vs velocity of a point on the treble bridge. The frequency of oscillation was 350 Hz. The slope of the graph is the driving-point impedance, which is 675±17 acoustic ohms.

the force applied to the soundboard, a force-dependent resistor was placed on the bridge at the point where the shaker touched it. An operational amplifier produced a voltage that was linearly proportional to the force exerted on the resistor.

Since the soundboard was driven harmonically at an angular frequency ω_0, the velocity of the driving point was given by

$$v = A_0\omega_0 \cos \omega_0 t, \tag{17}$$

where A_0 is the amplitude of the displacement. As the force was increased from zero, a black spot appeared at the position of the driving point when observed through the interferometer each time Eq. (14) was satisfied. This spot migrated into a ring as the driving amplitude was increased, and eventually a new spot appeared at the driving point. This process progressed until the limit of the driver was reached.

To determine the driving-point impedance, the force was measured each time the interferometry indicated that Eq. (14) was satisfied (i.e., a black spot appeared at the driving point as viewed through the interferometer). The force was plotted as a function of velocity, which was determined using Eq. (17), with A_0 determined from Eq. (14). The driving-point impedance is the slope of this graph, which was found by performing a linear regression. A typical graph of the force versus the velocity of the fully assembled soundboard is shown in Fig. 10. Each point is the result of three independent measurements of the displacement at the driving frequency. The uncertainty in the driving-point impedance is taken to be the uncertainty in the slope.

The advantages of this measurement system stem from the fact that the velocity is measured both directly and remotely. The remote measurement alleviates problems associated with the mass of an impedance head affecting the measurement of acceleration. Likewise, there are no resonances of the measuring system that can affect the results. Finally, this method provides an estimate of the uncertainty of the measurement, which is taken to be the uncertainty in the slope of the plot.

Before concluding this section, we note that the data presented here indicate that the approximations made in Sec. II are valid. Specifically, the linearity of the data used to determine the driving-point impedance, such as the example shown in Fig. 10, can only be possible if indeed the lines of equal displacement occur at the points determined by Eq. (14). Therefore, in the experimental arrangement described here, the assumptions that the functional form of Eq. (10) describes the movement of the soundboard and that $\epsilon \ll 1$ are valid. However, should an instance occur where these data are not linear, the value of ϵ could be determined simply by fitting the data to the equation $J_\epsilon(\xi)=0$, using ϵ as a fitting parameter.

V. CONCLUSIONS

We have described a method of electronic speckle pattern interferometry that not only works with moderate decorrelation of the speckle pattern, but demands it. We have shown theoretically and experimentally that this arrangement can be used to determine the deflection shapes of an object that is normally too unstable to observe interferometrically, and applied it to the study of a piano soundboard *in situ*.

Using this interferometer we have investigated the dynamics of the soundboard of a piano and have compared the results to a simple closed-form theory, as well as a finite-element model. Comparison of the deflection shapes of the piano to those predicted by these models demonstrates that the pressure exerted by the strings on the soundboard can make significant changes in mode shapes and resonant frequencies. The presence of this pressure has a significant effect on the lowest modes, but appears not to be important in determining the shapes and frequencies of the higher modes.

We have also shown that this interferometer can be used to determine resonance curves and driving-point impedance. We have presented the resonance curves for the lowest three modes of a soundboard and shown that they do not overlap significantly.

We close by noting that the applications of this interferometric technique are not restricted only to the investigation of piano soundboards. Harmonic vibrations of any unstable object that meets the requirements outlined in Sec. II can be observed using this technique. Additionally, the theory can be applied outside of the approximations if the value of ϵ in Eq. (10) is known.

ACKNOWLEDGMENTS

The authors thank B. Neller and M. Peterson for their generous support of this work. D. Parker assisted in the development of some of the computer models.

[1]K. Wogram, "Acoustical research on pianos. I. Vibrational characteristics of the soundboard," Das Musikinstrument **24**, 694–702 (1981).
[2]K. Wogram, "The strings and the soundboard," in *The Acoustics of the Piano*, edited by A. Askenfelt (Royal Swedish Academy of Music, Stockholm, 1990).
[3]H. A. Conklin, "Design and tone in the mechanoacoustic piano. II. Piano structure," J. Acoust. Soc. Am. **100**, 695–707 (1996).
[4]N. Giordano, "Simple model of a piano soundboard," J. Acoust. Soc. Am. **102**, 1159–1168 (1997).
[5]N. Giordano, "Sound production by a vibrating piano soundboard: Experi-

ment," J. Acoust. Soc. Am. **104**, 1648–1653 (1998).
[6]N. Giordano, "Mechanical impedance of a piano soundboard," J. Acoust. Soc. Am. **103**, 2128–2133 (1998).
[7]H. Suzuki, "Vibration and sound radiation of a piano soundboard," J. Acoust. Soc. Am. **80**, 1573–1582 (1986).
[8]J. Berthaut, M. Ichchou, and L. Jézéquel, "Piano soundboard: Structural behavior, numerical and experimental study in the modal range," Appl. Acoust. **64**, 1113–1136 (2003).
[9]J. Skala, "The piano soundboard behavior in relation with its mechanical admittance," in *Proc. of Stockholm Music Acoustics Conference (SMAC 03)*, 6–9 August, Royal Institute of Technology (KTH) (Royal Institute of Technology, Stockholm, 2003), pp. 171–174.
[10]E. J. Hansen, I. Bork, and T. D. Rossing, "Relating the radiated piano sound field to the vibrational modes of the soundboard," J. Acoust. Soc. Am. **114**, 2383 (2003).
[11]J. Kindel and I. Wang, "Modal analysis and finite element analysis of a piano soundboard," in *Proc. of 5th International Modal Analysis Conference*, 6–9 April, Imperial College of Science and Technology, London (Union College, Schenectady, NY, 1987), pp. 1545–1549.
[12]T. R. Moore, "A simple design for an electronic speckle pattern interferometer," Am. J. Phys. **72**, 1380 (2004).
[13]T. R. Moore, "Erratum: A simple design for an electronic speckle pattern interferometer," Am. J. Phys. **73**, 189 (2004).
[14]M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, NY, 1972).
[15]A. Leissa, *Vibrations of Plates* (Acoustical Society of America, Melville, NY, 1993).
[16]V. V. Bolotin, B. P. Makarov, G. V. Mishenkov, and Y. Y. Shveiko, "Asymptotic method of investigating the natural frequency spectrum of elastic plates," Raschetna Prochnost, Mashgiz **6**, 231–253 (1960).
[17]*Wood Handbook* (University Press of the Pacific, Honolulu, HI, 1974).

Contribution to harmonic balance calculations of self-sustained periodic oscillations with focus on single-reed instruments

Snorre Farner,[a] Christophe Vergez, Jean Kergomard, and Aude Lizée
Laboratoire de Mécanique et d'Acoustique, CNRS UPR 7051, 31 chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

(Received 9 March 2005; revised 19 December 2005; accepted 19 December 2005)

The harmonic balance method (HBM) was originally developed for finding periodic solutions of electronical and mechanical systems under a periodic force, but has been adapted to self-sustained musical instruments. Unlike time-domain methods, this frequency-domain method does not capture transients and so is not adapted for sound synthesis. However, its independence of time makes it very useful for studying any periodic solution, whether stable or unstable, without care of particular initial conditions in time. A computer program for solving general problems involving nonlinearly coupled exciter and resonator, HARMBAL, has been developed based on the HBM. The method as well as convergence improvements and continuation facilities are thoroughly presented and discussed in the present paper. Applications of the method are demonstrated, especially on problems with severe difficulties of convergence: the Helmholtz motion (square signals) of single-reed instruments when no losses are taken into account, the reed being modeled as a simple spring.
 [DOI: 10.1121/1.2166602]

PACS number(s): 43.75.Pq, 43.58.Ta [NHF] Pages: 1794–1804

I. INTRODUCTION

Since Helmholtz,[1] it has become natural to describe a self-sustained musical instrument as an exciter coupled to a resonator ["self-sustained" is a term indicating oscillation driven by a constant energy input.] More recently, McIntyre *et al.*[2] have highlighted that simple models are able to describe the main functioning of most self-sustained musical instruments. These models rely on few equations whose implementation is not CPU-demanding, mainly because the nonlinearity is spatially localized in an area small compared to the wavelength. This makes them well adapted for real-time computation (including both transient and steady states). These models are particularly popular in the framework of sound synthesis.

On the other hand, calculation in the frequency domain is suitable for determining periodic solutions of the model (the values of the harmonics as well as the playing frequency) for a given set of parameters. Such information can be provided by an iterative method called the harmonic balance method (HBM). Though the name "harmonic balance" seems to date back to 1936,[3] the method was popularized nearly 40 years ago for electrical and mechanical engineering purposes, first for forced vibrations,[4] later for auto-oscillating systems.[5] The modern version was presented rather shortly after by Nakhla and Vlach.[6] In 1978, Schumacher was the first to use the HBM for musical acoustics purposes with a focus on the clarinet.[7] However, in this paper, the playing frequency is not determined by the HBM. This shortcoming is the major improvement brought 11 years later by Gilbert *et al.*[8] who proposed a full study of the clarinet including the playing frequency as an unknown of the problem.

The fact that the HBM can only calculate periodic solutions may seem a drawback. Certainly, transients such as the attack are impossible to calculate, and the periodic result is boring to listen to and does not represent the musicality of the instrument. Therefore, the HBM is definitely not intended for sound synthesis. Nevertheless, self-sustained musical instruments are usually used to generate harmonic sounds, which are periodic by definition. The HBM is thus very useful for investigating the behavior of a physical model of an instrument, depending on its parameter values. This is possible for both stable and unstable solutions, without care of particular initial conditions, which are necessary in the time domain. Moreover, HBM results can be compared to approximate analytical calculations [like the variable truncation method (VTM)],[9] in order to check the validity of the approximate model considered.

The present paper is based on the work of Gilbert *et al.*[8] Our main contributions are an extension of the diversity of equations managed, improved convergence of the method, introduction of basic continuation facilities, and from a practical point of view, faster calculations. In order to test the convergence properties, especially when the number of harmonics increases, we treat a few extreme cases where exact solutions exist in the form of square (or "rectangular" signals) when losses are ignored. The solutions of such simple models of self-sustained instruments are known to be the so-called Helmholtz motion.

Recently, some of us published a paper discussing the different elements of related clarinet models by using the same software. The influence of the shape of the nonlinear function and of several parameters, such as the reed dynamics or the loss parameter, was studied in the context of cy-

[a]Currently at Department of Electronics and Telecommunications, NTNU, O. S. Bragstads pl. 2, 7491 Trondheim, Norway.

lindrical instruments.[10] In contrast, the model used in the present paper is essentially simple, with neither losses nor reed dynamics. This results in square or rectangular signals, corresponding to instruments with cylindrical and stepped-cone bores, respectively.

While the main idea was already described by Gilbert *et al.*,[8] Sec. II details the principle of the HBM, in particular the discretization of the problem, both in time and frequency.

Section III is devoted to the various contributions of the current work, which are applied in a computer program called HARMBAL.[11] The framework is defined to include models with three equations: two linear differential equations, written in the frequency domain, and a nonlinear coupling equation in the time domain (see Sec. III A). As usual in the HBM, this system of three equations is solved iteratively. The solving method chosen (Newton–Raphson, Sec. II B) has been investigated and its convergence has been improved through a backtracking scheme (Secs. III C and III D).

To illustrate the advantages of the HBM and the improvements, a few case studies were performed and are presented in Sec. IV. They are based on a classical model of single-reed instruments which is presented in Sec. IV A. In Secs. IV B and further, simplifications and variations of the model are introduced so that the results could be compared to analytical calculations, both for cylindrical and stepped-cone bores. Finally, the full model is compared to time-domain simulations. This also shows the modularity of HARMBAL. The comparison is achieved through the investigation of bifurcation diagrams as the dimensionless blowing pressure is altered. The derivation of a branch of solution is obtained thanks to basic continuation with an auto-adaptative parameter step.

Finally, various questions are tackled through practical experience from using HARMBAL. Section V discusses multiplicity of solutions and poor robustness in the frequency estimation.

II. NUMERICAL METHOD

A. The harmonic balance method

The harmonic balance method is a numerical method to calculate the steady-state spectrum of periodic solutions of a nonlinear dynamical system. The following provides a detailed and general description of the method for a nonlinearly coupled exciter-resonator system.

Let $X(\omega_k)$, $k=0,\ldots,N_t-1$ be the discrete Fourier transform (DFT) of one period $x(t)$, $0\leqslant t<T$, of a T-periodic solution of a mathematical system to be defined, where $\omega_k=2\pi k/N_t$. $X(\omega_k)$ will have a number of complex components N_t, which depends on the sampling frequency $f_s=1/T_s$ with which we discretize $x(t)$ into $N_t=T/T_s$ equidistant samples. Each (Fourier) component contains the amplitude and the phase of the corresponding harmonic of the signal. Note that the sampling frequency $f_s=N_t f_p$ is automatically adjusted to the current playing frequency f_p so that we always consider one period of the oscillation while keeping N_t constant. Note also that N_t should be sufficiently large to avoid aliasing. Moreover, if N_t is chosen as a power of 2, the fast Fourier transform (FFT) may be used. Assuming that $N_p<N_t/2$ harmonics are sufficient to describe the solution, we define $\mathbf{X}\in\mathbb{R}^{2N_p+2}$ (i.e., a vector of $2N_p+2$ real components) as the first N_p+1 real components (denoted by $\Re$) of $X(\omega_k)$ followed by their imaginary counterparts ($\Im$)

$$\mathbf{X}=[\Re(X(\omega_0)),\ldots,\Re(X(\omega_{N_p})),\ \Im(X(\omega_0)),\ldots,\Im(X(\omega_{N_p}))]. \tag{1}$$

Note that the components X_0 and X_{N_p+1} are the real and imaginary dc components, respectively (and that X_{N_p+1} is always zero). Our mathematical system can thus be defined by the nonlinear function $F:\mathbb{R}^{2N_p+3}\rightarrow\mathbb{R}^{2N_p+2}$

$$\mathbf{X}=\mathbf{F}(\mathbf{X},f_p). \tag{2}$$

Until now, the playing frequency has silently been assumed to be a known quantity. In autonomous systems, however, the frequency is an additional unknown, so that the N_p-harmonic solution sought is defined by $2N_p+3$ unknowns linked through the $2N_p+2$ equations (2). However, it is well known that as $\mathbf{X}$ is a periodic solution of a dynamical system, and $\mathbf{X}'$ deduced from $\mathbf{X}$ by a phase rotation (i.e., a shift in the time domain) is also a solution. Thus, an additional constraint has to be added in order to select a single periodic solution among the infinity of phase-rotated solutions. A common choice (see Ref. 8) is to consider the solution for which the first harmonic is real (i.e., its imaginary part, X_{N_p+2}, is zero). This additional constraint decreases the number of unknowns to $2N_p+2$ for an N_p-harmonic periodic solution. Thus, we get $\mathbf{F}:\mathbb{R}^{2N+2}\rightarrow\mathbb{R}^{2N+2}$, and it is now possible to find periodic solutions, if they exist.

Finally, a simple way of avoiding trivial solutions to Eq. (2) is to look for roots of the function $\mathbf{G}:\mathbb{R}^{2N_p+2}\rightarrow\mathbb{R}^{2N_p+2}$, defined by

$$\mathbf{G}(\mathbf{X},f_p)=\frac{\mathbf{X}-\mathbf{F}(\mathbf{X},f_p)}{X_1}, \tag{3}$$

i.e., $\mathbf{G}(\mathbf{X},f_p)=0$. This equation is usually solved numerically through an iteration process, for instance by the Newton–Raphson method as in our case. How to handle the playing frequency f_p will be discussed in the following section.

B. Iteration by Newton–Raphson

The equation $\mathbf{G}(\mathbf{X},f)=0$, $\mathbf{G}$ being defined by Eq. (3), is nonlinear and usually has no analytical solution (for readability we leave out the index p on the playing frequency until the end of Sec. III). The Newton–Raphson method is a multidimensional extension of the well-known Newton's method, both of which are available in many text books on calculus, e.g., Ref. 12, Sec. 9.6. This is the method used in the program HARMBAL (see Sec. III), although it had to be refined with a backtracking procedure to improve its convergence, as discussed in Sec. III D.

In our $2N_p+2$-dimensional case, we have a vector problem: we search $(\mathbf{X},f)$ for which $\mathbf{G}(\mathbf{X},f)=0$. As highlighted by Gilbert *et al.*,[8] the playing frequency is unknown and must be included in the root-finding process

$$(\mathbf{X}^{i+1},f^{i+1}) = (\mathbf{X}^i,f^i) - (\mathbf{J}_G^i)^{-1}\cdot\mathbf{G}(\mathbf{X}^i,f^i), \tag{4}$$

where $\mathbf{J}_G^i \triangleq \nabla G(\mathbf{X}^i,f^i)$ is the *Jacobian* matrix of $\mathbf{G}$ at $(\mathbf{X}^i,f^i)$. (The symbol $\triangleq$ indicates that the relation is a definition.) Note that all derivatives by X_{N_p+2}, which was chosen to be zero, are ignored. The column N_p+2 in the Jacobian is thus replaced by the derivatives with respect to the playing frequency f. $\mathbf{J}_G^i$ is thus a $(2N_p+2)$-square matrix. This means that line number N_p+2 in Eq. (4) gives the new frequency f instead of X_{N_p+2}. We define the *Newton step* $\Delta\mathbf{X}=\mathbf{X}^{i+1}-\mathbf{X}^i$ (where $\Delta f=f^{i+1}-f^i$ replaces ΔX_{N_p+2}), which follows the local steepest descent direction.

The Jacobian may be found analytically if $\mathbf{G}$ is given analytically, but it is usually sufficient to use the first-order approximation

$$J_{jk} = \frac{\partial G_j}{\partial X_k} \simeq \frac{G_j(\mathbf{X}+\delta\mathbf{X}_k,f) - G_j(\mathbf{X},f)}{\delta X}, \tag{5}$$

except for $k=N_p+2$, in which case we use

$$J_{j,N+2} = \frac{\partial G_j}{\partial f} \simeq \frac{G_j(\mathbf{X},f+\delta f) - G_j(\mathbf{X},f)}{\delta f}. \tag{6}$$

The components of $\delta\mathbf{X}_k$ are zero except for the kth one, which is the tiny perturbation $\delta X=10^{-5}|X_1|$ or $\delta f=10^{-5}f$. The iteration has converged when $|\mathbf{G}^i| \triangleq |\mathbf{G}(\mathbf{X}^i,f^i)| < \varepsilon$. We found $\varepsilon=10^{-5}$ to be a good compromise between computation time and solution accuracy.

III. IMPLEMENTATION AND HARMBAL

A. Equations for self-sustained musical instruments

Though, to the authors' knowledge, the harmonic balance method in the context of musical acoustics with unknown playing frequency has only been applied to study models of clarinet-like instruments, it should be possible to consider many different classes of self-sustained instruments. It is well accepted that sound production by a musical instrument results from the interaction between an exciter and a resonator through a nonlinear coupling. Moreover, in most playing conditions, linear modeling of both the exciter and the resonator is a good approximation.

Therefore, within these hypotheses, any self-sustained musical instrument could be modeled by the following three equations:

$$\begin{cases} Z_e(\omega)X_e(\omega) = X_c(\omega) & \text{(7a)} \\ X_c(\omega) = Z_r(\omega)X_r(\omega) & \text{(7b)} \\ \mathcal{F}[x_c(t),x_e(t),x_r(t)] = 0, & \text{(7c)} \end{cases}$$

where Z_e and Z_r are the input impedances of the exciter and the resonator, respectively, and X_e and X_r are the spectra describing the dynamics of the exciter and the resonator during the steady state (periodicity assumption). X_c is the spectrum of the coupling variable. All these quantities, and thus Eqs. (7a) and (7b), are defined in the Fourier domain. Equation (7c) is written in the time domain, where $\mathcal{F}$ is a nonlinear functional of x_c, x_e, and x_r, which are the inverse Fourier transforms of X_c, X_e, and X_r, respectively. We apply the discretization as described in Sec. II A, implying that Eqs. (7a) and (7b) become vector equations where the impedances must be written as real $(2N_p+2)\times(2N_p+2)$-matrices to accommodate the rules of complex multiplication

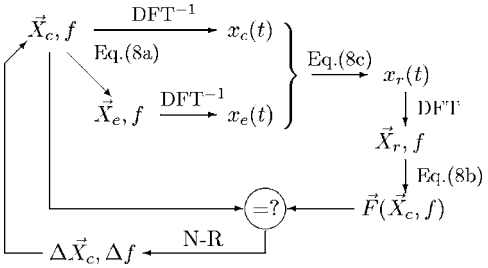


FIG. 1. The iteration loop of the harmonic balance method for a musical instrument (notations defined in the text).

$$Z(f) = \begin{pmatrix} \Re[\tilde{Z}(f)] & -\Im[\tilde{Z}(f)] \\ \Im[\tilde{Z}(f)] & \Re[\tilde{Z}(f)] \end{pmatrix} \tag{8}$$

where

$$\tilde{Z}(f) = \begin{pmatrix} Z(0) & 0 & \cdots & 0 \\ 0 & Z(\omega_1) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & Z(\omega_{N_p}) \end{pmatrix} \tag{9}$$

is complex, and $\Re(\tilde{Z})$ and $\Im(\tilde{Z})$ are the real and imaginary components of $\tilde{Z}$. The system (7) is solved iteratively by HARMBAL according to the scheme illustrated in Fig. 1.

In HARMBAL, these equations are easily defined by writing small, new C functions. Only superficial knowledge of the C language is necessary to do this.

Three cases related to models of single-reed instruments with cylindrical or stepped-conical bores are studied in particular in Sec. IV in order to validate the code and to illustrate the modularity of HARMBAL and the HBM.

B. Practical characteristics of HARMBAL

Fast calculation, good portability, and independence of commercial software are easily achieved by programming in C, whose compiler is freely available for most computer platforms. It is, however, somewhat difficult to combine portability with easy usage, because an intuitive usage normally means a graphical and interactive user interface, while the handling of graphics varies greatly between the different platforms.

We have chosen to write HARMBAL with a nongraphical and noninteractive user interface. (The term "noninteractive" means that the user has no influence on the program while it is running.) The major advantage of this is that independent user interfaces may be further developed depending on need.

Our concept is to save both the parameters and the solution in a single file. This file also serves as input to HARMBAL while individual parameters can be changed through start-up arguments. The solution provided by the file works as the initial condition for the harmonic balance method. Thus, the lack of a simple user interface is compensated by a simple way of reusing an existing solution to solve the sys-

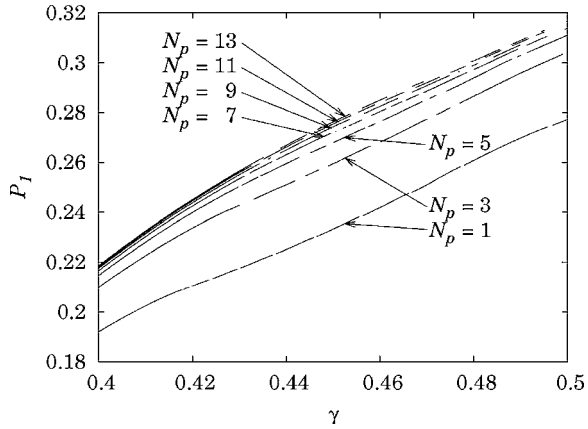


FIG. 2. Solution holes: first pressure harmonic P_1 versus blowing pressure γ for different N_p with N_t=128, ζ=0.5, and η=10^{-3}. (Even N_p give the same as N_p-1.) Equations and parameters are defined in Sec. IV.

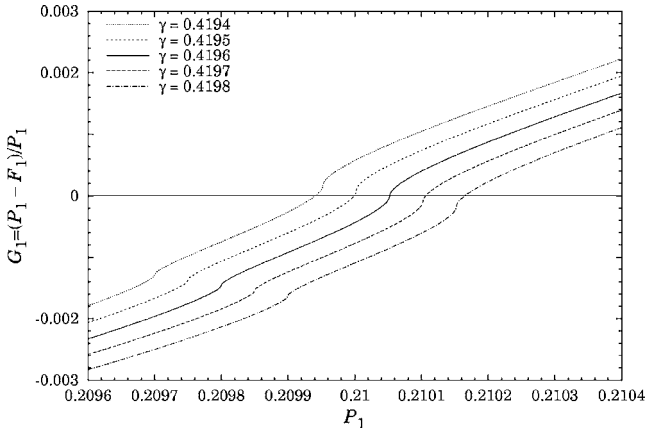


FIG. 3. G_1 as a function of P_1 around the solution G_1=0 for various γ around a hole at $\gamma\simeq$0.4196. N_t=128 and N_p=1.

tem for a slightly different set of parameters. Solutions for a range of a parameter values may thereby be calculated by changing the parameter stepwise and providing the previous solution as an initial condition for the next run. The Perl script *hbmap* provides such zeroth-order continuation facilities. This procedure may also be used when searching for a solution where it is difficult to provide a sufficiently good initial condition, for instance by successively increasing N_p when wanting many harmonics.

C. Convergence of Newton–Raphson

When employing the method in its standard version to determine the solution of the system at a given set of parameters, we have found that it is impossible to find a solution at particular combinations of parameter values. Indeed, for the clarinet model of Sec. IV B 1, no convergence was obtained for particular values of the parameter γ (the dimensionless blowing pressure) and its neighborhood. This is seen as discontinuities, or *holes*, in the curves in Fig. 2 (see Sec. IV for the underlying equations and parameters). Note that the solutions seem to go continuously through this hole and that the positions of the holes and their extent vary with the number of harmonics N_p taken into account. The curves were obtained by using *hbmap* to calculate a quasicontinuity of solutions descending from γ=0.5 in steps of 10^{-4} and drawing a line between them, except across γ values where solution failed. In these holes, the Newton–Raphson method did not converge either by alternating between two values of $\mathbf{P}$ (i.e., $\mathbf{X}_c$) or by starting to diverge.

To study the problem, we simplified the system to a one-dimensional problem by setting N_p=1, thus leaving P_1 as the only nonzero value. G_1 thus became the only contributor to $|\mathbf{G}|$, and a simple graph of G_1 around the solution G_1=0 could illustrate the problem, as shown in Fig. 3 for several values of γ. We see that the curve of $G_1(P_1)$ has inflection points (visible as "soft steps" on the curve) at rather regular distances. At the center of a convergence hole, in this case at γ=0.4196, an inflection point is located at the intersection with the horizontal axis. This is a school example of a situation where Newton's method (the one-dimensional limit of the Newton–Raphson method) does not converge because the Newton step ΔP_1 brings us alternatingly from one side of the solution to the other, but not closer.

In fact, the existence of inflection points is linked with the digital sampling of the continuous signal. If the sampling rate is increased, i.e., if N_t is increased, the steps become smaller but occur more frequently, as shown for N_t=32, 128, and 1024 in Figs. 4(a)–4(c). The derivative dG_1/dP_1 is included in the figures to quantify the importance of the steps. According to Figs. 4(a)–4(c), it seems reasonable to increase N_t to avoid convergence problems. However, this would significantly increase the computational cost. Another solution is therefore suggested in the following.

D. Backtracking

When the Newton–Raphson scheme fails to converge, it often happens because the Newton step $\Delta\mathbf{X}$ leads to a point

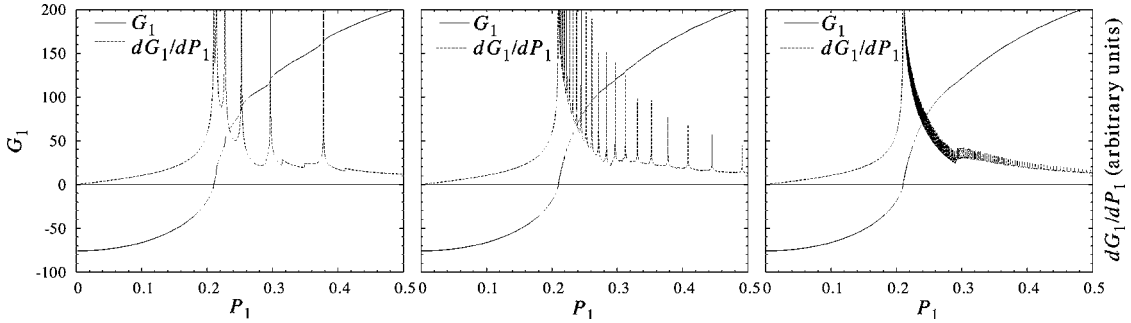


FIG. 4. The effect of sampling rate on the "smoothness" of $G_1(P_1)$: (a) N_t=32; (b) 128; and (c) 1024. The derivative dG_1/dP_1 exhibits the "roughness."

TABLE I. Description of variables and constants. The dimensionless variables and parameters used in Sec. IV are derived from dimensional variables [denoted by a hat (∧)] as follows: $x_e=x=\hat{y}/H+\gamma/K$, $x_c=p=\hat{p}/p_M$, $x_r=u=\hat{u}\rho c/Sp_M$, $\omega=\hat{\omega}/f_r$, $\gamma=p_m/p_M$, and $\zeta=(Hw/S)\sqrt{2\rho c^2/p_M}$. The dimensionless time is $t=\hat{t}\omega_r$. The coefficients of the reed equation are $M=K\omega_r^2/\omega_e^2$, $R=Kg_e\omega_r/\omega_e^2$, and $K=\mu_e H\omega_e^2/p_M$. Note that index e denotes exciter and r resonator. SI units in parentheses.

$\hat{y}$	displacement of reed from equilibrium (m)
$\hat{p}$	pressure in reed opening (Pa)
$\hat{u}$	volume flow though reed opening (m^3/s)
$\hat{f}_r$	first mode frequency of pipe (Hz)
f_p	playing frequency (Hz)
p_m	mouth pressure/blowing pressure (Pa)
p_M	pressure p_m that closes reed opening (Pa)
H	equilibrium height of reed opening (m)
w	width of reed opening (m)
S	cross section of the pipe (m)
g_e	damping of the reed (s^{-1})
μ_e	specific mass of the reed (kg/m^2)
ω_e	first mode angular frequency of the reed (s^{-1})
ω_r	first mode angular frequency of the pipe ($2\pi f_r$; s^{-1})
Z_c	characteristic input impedance of the pipe ($\rho c/S$)
η	dimensionless loss parameter for the pipe
ρ	density of air (kg/m^3)
c	sound speed in air (m/s)

where $|\mathbf{G}(\mathbf{X},f)|$ is larger than in the previous step. However, acknowledging that the Newton step points in the direction of the steepest descent, there must be a point along $\Delta\mathbf{X}$ where $|\mathbf{G}(\mathbf{X},f)|$ is smaller than in the previous iteration of the HBM. A backtracking algorithm described in Sec. 9.7 of Ref. 12 (see the Appendix) solves the problem elegantly by shortening the Newton step to $\lambda\Delta\mathbf{X}$ for $0<\lambda<1$.

IV. CASE STUDIES

A. Equations for the clarinet

The three equations (7a)–(7c) may be constructed by physical modeling. In the case of the clarinet, a common simple model can be constructed by a reed equation nonlinearly coupled to a pipe equation by the Bernoulli equation.[7,10,13–15] The equations are summarized below in dimensionless form with the variables and constants defined in Table I. The corresponding mouthpiece is illustrated in Fig. 5.

The exciter is an oscillating reed which may be modeled as a spring with mass and damping

$$M\ddot{x}+R\dot{x}+Kx=p, \tag{10}$$

where the dimensionless reed displacement x (dots denoting time derivative) is the exciter variable x_e and the pressure p in the mouthpiece is the coupling variable x_c. The coefficients $M=\omega_r^2/\omega_e^2$, $R=g_e\omega_r/\omega_e^2$, and $K=1$ are the dimensionless mass, damping, and spring constant of the reed, where g_e is the reed damping and ω_r and ω_e are the first mode angular frequencies of the resonator and the exciter, respectively. $K=1$ because the reed closes when blowing at the maximum blowing pressure in the static regime ($p=0$). The exciter impedance in Eq. (7a) thus becomes

$$Z_e(\omega)=K-M\left(\frac{\omega}{2\pi}\right)^2+iR\left(\frac{\omega}{2\pi}\right), \tag{11}$$

where the dimensionless angular frequency ω is 2π at the first resonance of the pipe.

The cylindrical quarter-wave resonator is simply described by the dimensionless input impedance to be used in Eq. (7b)

$$Z_r(\omega)=i\tan\left(\frac{\omega}{4}-i\alpha(\omega)\right), \tag{12}$$

where $\alpha(\omega)\triangleq\psi\eta\sqrt{\omega/2\pi}$ is the dissipation in the tube, $\psi\simeq 1.3$, and η being the dimensionless loss parameter. Here, η depends on the tube length and is typically 0.02 for a normal clarinet with all holes closed. We set η to almost zero in our lossless calculations.

Finally, the nonlinear coupling equation (7c) is given by relating the volume flow u of air through the reed opening, i.e., the coupling variable x_c, to x and p by the Bernoulli equation. Taking into account high blowing pressures γ, the dimensionless version becomes[14]

$$u(p,x)=\zeta(1+x-\gamma)\sqrt{|\gamma-p|}\,\mathrm{sign}(\gamma-p), \tag{13}$$

as long as $x>\gamma-1$, and $u=0$ otherwise because the reed closes the mouthpiece, i.e., the reed "beats." ζ is a dimensionless embouchure parameter roughly describing the mouthpiece and the position of the player's lips. Only the nonbeating-reed regime is considered in this study.

Disregarding the reed dynamics, i.e., considering the reed mass and damping negligible, we have $x=p$ instead of Eq. (10). The exciter impedance thus simplifies to $Z_e=1$ and the coupling equation to

$$u(p)=\zeta(1+p-\gamma)\sqrt{|\gamma-p|}\,\mathrm{sign}(\gamma-p) \tag{14}$$

for $p>\gamma-1$, and, as before, $u=0$ otherwise.

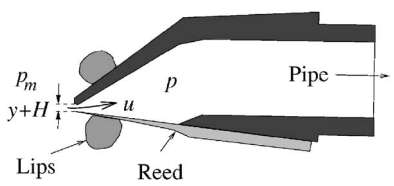


FIG. 5. Illustration of the mouthpiece. The mouth pressure p_m becomes γ in dimensionless quantities, and the instantaneous reed opening $y+H$ becomes $x_e-\gamma+1$ (see Table I).

B. Verification of method and models

In the following we want to verify that the HBM (and its implementation in HARMBAL) gives correct results. By using very low losses in the resonator (small η) and disregarding the reed dynamics, we can compare the results of the HBM with analytical results. Moreover, we can compare our results with numerical results from real-time synthesis. For this we need to include mass and damping of the exciter. These case studies also illustrate the modularity of HARMBAL as both the exciter and the resonator are changed. An example of changing the coupling equation is shown by Fritz *et al.*[10]

1. Helmholtz oscillation for cylindrical tubes

To compare the HBM results with analytical results, we assume a nondissipative air column, i.e., setting $\eta=0$ and

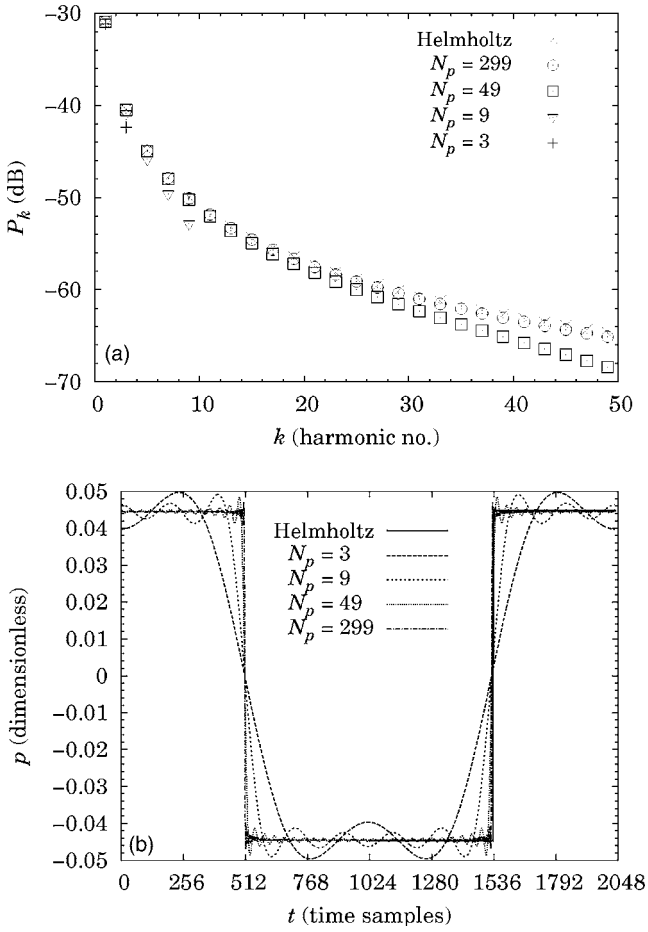


FIG. 6. The Helmholtz solution, Eq. (15), compared with the HBM truncated to 3, 9, 49, and 299 harmonics close to the oscillation threshold ($\gamma=0.334$, $\zeta=0.5$, $\eta=10^{-5}$) in (a) frequency and (b) time domain.

thus $\alpha=0$ in Eq. (12). Furthermore, we neglect reed dynamics and thus use $Z_e=1$ and Eq. (14). The resulting square-wave amplitude (the Helmholtz motion)[1] may be found by solving $u(p)=u(-p)$, which results from the fact that the internal pressure $p(t)$ and the power $p(t)u(t)$ averaged over a period are zero according to the lossless hypothesis.[14] This leads to a square-wave oscillation with amplitude

$$p(\gamma)=\sqrt{-3\gamma^2+4\gamma-1}. \tag{15}$$

As shown in Fig. 6, the HBM solution close to the oscillation threshold shows very good convergence towards the Helmholtz motion as the number of harmonics increases. Note that the points do not match for higher harmonics, not even for 299 harmonics. Dissipation in the resonator ($\eta\neq 0$) causes higher harmonics to be damped more close to the threshold ($\gamma\simeq 1/3$) than for higher blowing pressures (as explained, e.g., in Ref. 9). The deviation from a square-wave signal is thus more noticeable close to the threshold, and as the HBM calculations imposed a nonzero dissipation, this is probably the reason for the small deviation in Fig. 6. Our calculations at $\gamma=0.4$ confirm that the HBM results for $N_p=299$ approach the Helmholtz motion further away from the threshold.

Figure 2 in Ref. 10 shows P_1 as a function of γ for various N_p for $\eta=0.02$, calculated by HARMBAL. A similar diagram may be made for the lossless case here, as shown in

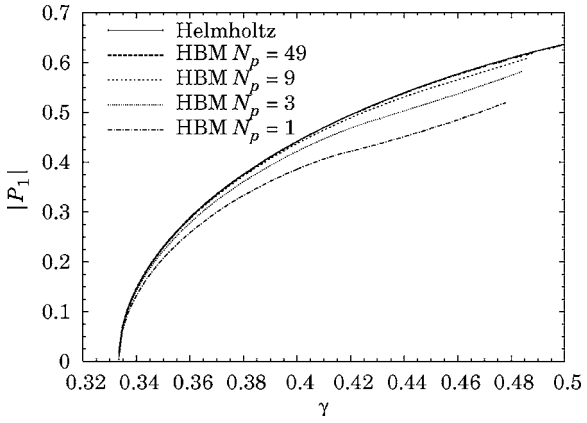


FIG. 7. Amplitude of first harmonic as the blowing pressure increases for the Helmholtz solution (15) and the HBM truncated to N_p harmonics ($\zeta=0.5$, $\eta=10^{-5}$)

Fig. 7. The oscillation threshold is reduced to $\gamma=1/3$, at which point the model experiences a direct Hopf bifurcation (which is known since the work of Grand *et al.*),[16] i.e., the oscillation starts continuously from zero as γ is increased beyond the threshold. This means that a single harmonic is enough to study the solution around the threshold. Far from the threshold, more harmonics have to be taken into account for P_1 to converge toward the Helmholtz solution. This is not obvious and contradictory to the hypothesis made for the VTM,[9] for example. Thus, HARMBAL appears as an interesting tool to evaluate the relevance of approximate methods at certain parameter values.

2. Helmholtz oscillation for a stepped conical tube

The saxophone works similarly to the clarinet, but the bore has a conical shape. In this section we compare HBM calculations with analytical results, so we simplify the cone by assuming that it consists of a sequence of N cylinders of length l and cross section $S_i=\frac{1}{2}i(i+1)S_1$ for $i=1,\ldots,N$ and $S_1=S$ being the cross section of the smallest cylinder (see Ref. 17). The total length of the instrument is thus $L=Nl$. The input impedance of such a *stepped cone* may be written as

$$Z_r(\omega)=\frac{2i}{\cot(\omega'/4-i\alpha(\omega'))+\cot(\omega' N/4-i\alpha(\omega' N))}, \tag{16}$$

where $\omega'\triangleq 2\omega/(N+1)$ so that $\omega=2\pi$ at the first resonance of this resonator. Equation (16) is used instead of Eq. (12), and the losses $\alpha(\omega)$ are zero in the analytic Helmholtz case and very small for the HBM calculations ($\eta=2\cdot 10^{-5}$, below which convergence became difficult).

Similarly to the cylinder case, the pressure amplitude of the ideal lossless stepped cone is calculated by solving $u(p)=u(-Np)$. Two solutions are possible [This result corrects Eq. (14) in Ref. 17]:

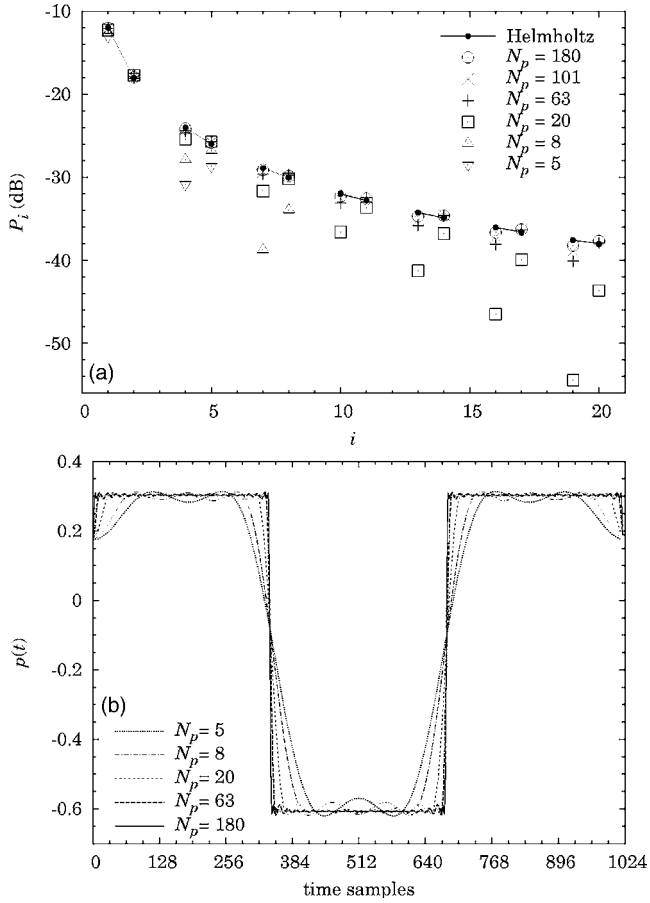


FIG. 8. Comparison between the standard Helmholtz motion of a stepped cone (N=2) and the HBM for various N_p at γ=0.31, ζ=0.2, and $\eta = 2\cdot 10^{-5}$. (a) The magnitude of the harmonics and (b) one oscillation period. N_t varies from 128 for N_p=5 to 1024 for N_p=180.

$$p^{\pm}(\gamma) = \frac{(N-1)(2-3\gamma)}{2(N^2-N+1)} \pm \frac{\sqrt{(N-1)^2+(N+1)^2(-3\gamma^2+4\gamma-1)}}{2(N^2-N+1)}, \qquad (17)$$

as long as $\gamma < 1/(N+1)$ for the standard Helmholtz motion (p^+) and $\gamma < N/(N+1)$ for the inverted one (p^-), which is unstable. Above these limits $p^+=\gamma$ and $p^-=-\gamma/N$. The magnitude of the first harmonic of a square or rectangular wave is then given by

$$P_1^{\pm}(\gamma) = \frac{\sin(\pi/(N+1))}{\pi/(N+1)} p^{\pm}(\gamma). \qquad (18)$$

For N=1, Eq. (17) reduces to Eq. (15). For higher N, the pressure wave becomes asymmetric.

We take the case N=2 and get

$$p^{\pm}(\gamma) = \tfrac{1}{6}(2-3\gamma \pm \sqrt{-27\gamma^2+36\gamma-8}). \qquad (19)$$

This result is compared with HBM calculations in Fig. 8 for γ=0.31. Theoretically, the spectrum of the Helmholtz solution, Fig. 8(a), shows that every third component is missing (actually zero) while the remaining components decrease in magnitude, thus forming the asymmetric pressure oscillation as shown in Fig. 8(b). The HBM, on the other hand, arrives at a solution where the first component

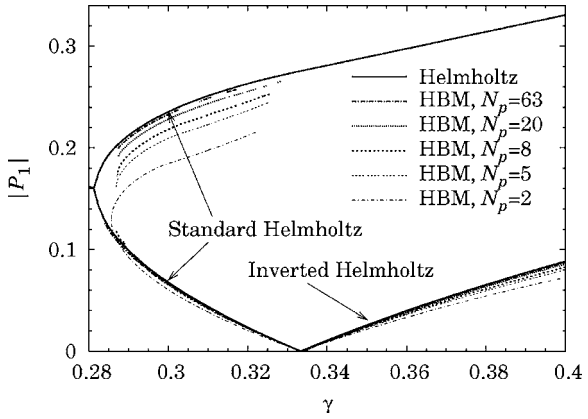


FIG. 9. Amplitude of first harmonic P_1 as a function of the blowing pressure γ for the Helmholtz solution (19) for two-stepped cone and the HBM truncated to 2, 5,…, and 63 harmonics, the last coinciding with Helmholtz (ζ =0.2, $\eta = 2\cdot 10^{-5}$). Only nonbeating regimes are shown.

in each pair of nonmissing components, i.e., components 1, 4, 7, etc., deviate more from the Helmholtz solution than the second component, i.e., 2, 5, 8, etc. This results in a *dip* at the middle of the long, positive part of the period [i.e., on both extreme times 0 and 1024 of the curve in Fig. 8(b)]. The same was observed for N=3 and N=4, where the long part of the period was divided by similar dips into three and four parts, respectively (not shown). The number of time samples, N_t, did not change this fact, but as Fig. 8 indicates, the dips gradually become narrower as the number of harmonics N_p increases. This indicates that the HBM approaches the Helmholtz solution as N_p approaches infinity.

A bifurcation diagram is plotted in Fig. 9, i.e., the amplitude of the first harmonic is plotted for different numbers of harmonics as a function of the blowing pressure γ. This was done by first finding the solution at γ=0.31; then, the script *hbmap* was used to make HARMBAL calculate a solution for each of many consecutive values of γ down to the oscillation threshold by using the previous solution as initial value. The procedure was repeated from the solution at γ =0.31 up to the point where the reed started to beat, i.e., where $p < \gamma - 1$ in Eq. (14). In practice, these curves are more difficult to obtain with *hbmap* than for the cylindrical bore, especially close to the subcritical oscillation threshold around γ=0.28, where computation was not possible at these low losses. More sophisticated continuation schemes should be considered to obtain complete curves.

The Helmholtz solution [Eq. (18) with N=2] is included with a solid line in Fig. 9. The lower part of the standard Helmholz branch and the branch of the inverted Helmholtz motion (see the figure) are unstable.[18]

Despite the uncompleted curves, the diagram shows that the model experiences a subcritical Hopf bifurcation, which agrees with the conclusion of Grand *et al.*[16] This means that a single-harmonic approximation is not enough to study the solution around this threshold, since the small-amplitude hypothesis with few harmonics close to the threshold does not hold. Convergence toward the Helmholtz motion is ensured as the number of harmonics N_p is increased. It can be noted that the beating threshold for the model with N_p harmonics

TABLE II. The values of M and R for three strengths of reed interaction. The bore parameters are $D=247$ ($f_r=103.4$ Hz), $a_1=0.899$, and $b_0=0.0946$ for sampling frequency $f_s=51\,100$ Hz

Reed	$\omega_e/2\pi$ Hz	q_e	M	R
Weak	10 000	0.1	$1.070\cdot10^{-4}$	$1.034\cdot10^{-3}$
Normal	2 500	0.2	$1.712\cdot10^{-3}$	$8.28\cdot10^{-3}$

depends on N_p but converges toward the Helmholtz threshold $\gamma=1/3$ (corresponding to the lossless, continuous system) as N_p is increased.

3. Validation with time-domain model

When adding mass and damping to the reed, and viscous losses or dispersion to the pipe, it becomes more difficult to find analytic solutions to compare with the HBM. As far as the playing frequency is concerned, this has been studied by Fritz *et al.*[10] by comparison with an approximate analytical formula. Here, we compare both the playing frequency and the amplitude of the first partial with numerical results obtained with a time-domain method. We use a newly developed (real-time) time-domain method (here called TDM) by Guillemain *et al.*[19] It is based on the same set of equations as presented in Sec. IV A except that the impedance of the bore is modified a little to be expressed as an infinite impulse response. In the Fourier domain it becomes

$$Z_r(\tilde{\omega}) = \frac{1 - a_1 e^{-i\tilde{\omega}} - b_0 e^{-i\tilde{\omega}D}}{1 - a_1 e^{-i\tilde{\omega}} + b_0 e^{-i\tilde{\omega}D}}, \tag{20}$$

where $\tilde{\omega}$ is the angular frequency normalized by the sampling frequency f_s, and the integer $D=\text{round}(f_s/2f_r)$ is the time delay in samples for the sound wave to propagate to the end of the bore and back. The constants a_1 and b_0 are to be adjusted so that the first two peaks of resonance have the same amplitude as the first two peaks of Eq. (12).

To express Eq. (20) using our terminology, we write that $\omega=\tilde{\omega}f_s/f_r$ and obtain

$$Z_r(\omega) = \frac{1 - a_1 e^{-i\omega f_r/f_s} - b_0 e^{-i\omega/2}}{1 - a_1 e^{-i\omega f_r/f_s} + b_0 e^{-i\omega/2}}. \tag{21}$$

Because the TDM does not work for zero or even for small values of the mass M and damping R, we include reed dynamics in the calculations and use a reed with weak interaction with the pipe resonance as well as one with close to normal reed impedance. The corresponding values for ω_e and $q_e \triangleq g_e/\omega_e$ are shown in Table II. Figure 10(a) shows the bifurcation diagram for two values of ζ and for weak and normal reed impedance, while Fig. 10(b) shows the corresponding variation in the dimensionless playing frequency f_p/f_r. The lines represent the continuous solutions of the HBM, and the symbols show a set of results derived from the steady-state part of the TDM signal. The TDM symbols fall well on the lines of the HBM, except for $\zeta=0.50$ when γ approaches 0.5. Then, the TDM experiences period doubling, i.e., two consecutive periods of the signal differ. At the same time, not being able to show subharmonics, the HBM shows signs of a beating reed, possibly a solution that is unstable and thus not attainable by time-domain methods.

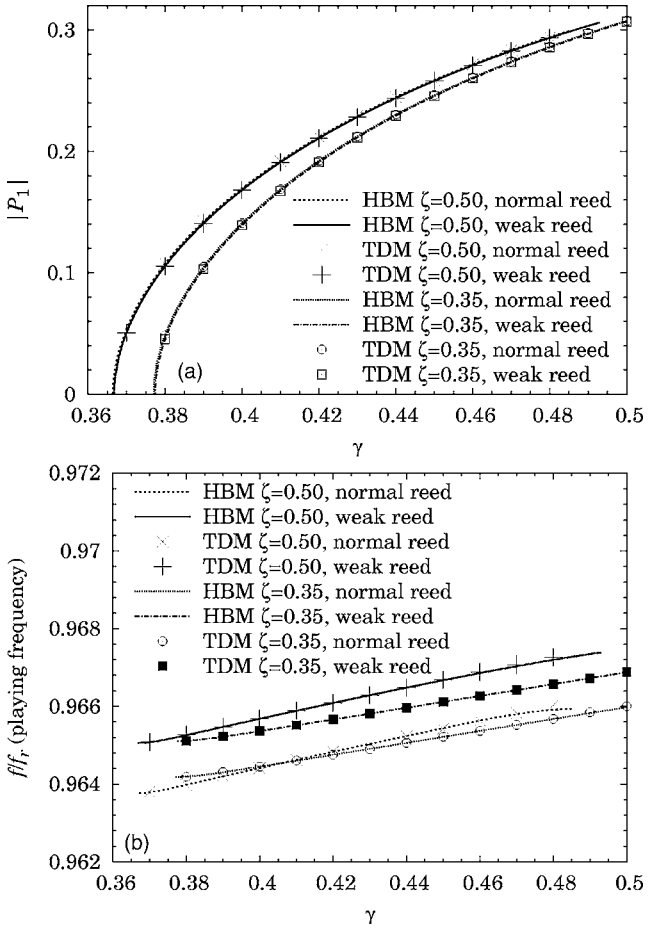


FIG. 10. Comparison between HBM and TDM of the amplitude of (a) the first harmonic P_1 and (b) the dimensionless playing frequency f_p/f_r as the blowing pressure γ increases for a clarinet-like system with viscous losses and weak and normal reed interaction. TDM values for $\zeta=0.50$ and $\gamma>0.48$ are omitted due to period doubling, as are the beating regimes of HBM calculations ($f_s=51100$ Hz, $N_t=512$, $f_r=103.4$ Hz, $N_p=15$).

Note that three points have to be verified before comparing results from the HBM and the TDM:

(1) The numerical scheme used in the TDM to approximate the time derivatives in the reed equation (10) requires discretization. Depending on the sampling frequency f_s, the peak of resonance of the reed deviates more or less from the one given by the continuous equation. For normal reed interaction ($\omega_e/2\pi=2500$ Hz), the deviation is negligible, but it may become significant in the case of weak reed interaction, where the peak is at 10 000 Hz. However, the fact that the reed and the bore interact weakly in the latter case implies that the exact position of the peak has little importance. Therefore, at the used sampling frequency, it was not necessary to compensate for the discretization in the TDM in the HBM calculations.

(2) Then, there should be agreement between the sampling frequency f_s in the TDM and the number of samples N_t per period in the HBM. Their relation is given by

$$N_t = \frac{f_s}{f_p}. \tag{22}$$

In order to have a sufficiently high sampling rate, we

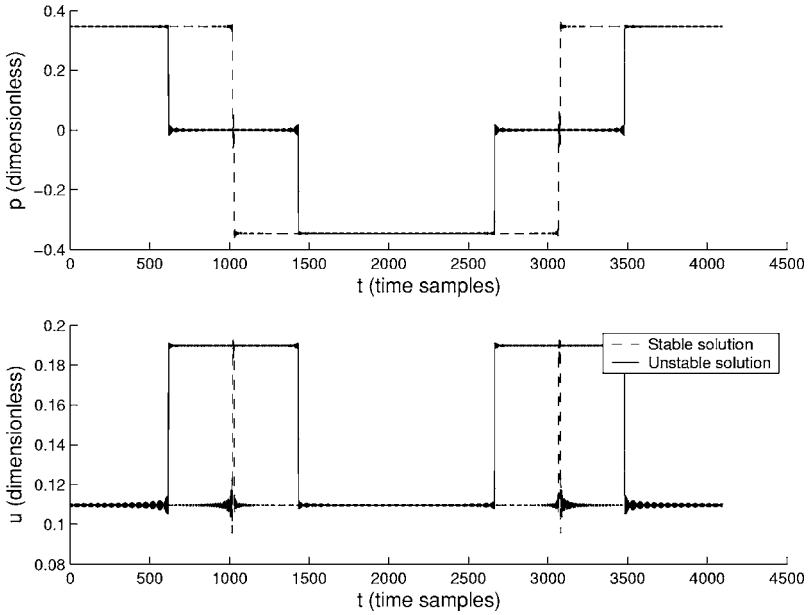


FIG. 11. The pressure and volume-flow wave form of the Helmholtz solution and a three-level sister solution calculated by the HBM employing the simple clarinet model with $\zeta=0.5$, $\gamma=0.4$, $N_p=2000$, $\eta=10^{-5}$.

have chosen $N_t=512$. The playing frequency f_p is plotted in Fig. 10(b), and we used an average $f_s=51\,100$ Hz for both the HBM and the TDM.

(3) Finally, it also seems necessary that N_p and N_t are chosen so that

$$N_p + 1 = \frac{N_t}{2}. \tag{23}$$

In practice, however, when comparing bifurcation diagrams of the first harmonic P_1, as in Fig. 10, rather low values of N_p give good results. Nevertheless, more harmonics are obviously needed to compare waveforms in the time domain, especially far from the oscillation threshold.

V. PRACTICAL EXPERIENCES

A. Multiple solutions

As we consider a nonlinear problem, we cannot anticipate the number of solutions. Therefore, it should not be surprising that it is possible to obtain multiple solutions for a given set of parameter values. When searching for a particular solution, this may be a practical problem. Fritz *et al.*[10] have discovered that some solutions seem to disappear when increasing the number of harmonics N_p, implying that solutions may arise from the truncation to a finite N_p. We have now discovered alternative solutions that persist even at very high N_p.

Let us illustrate this with the simple model of the clarinet used in Sec. IV B 1, where the reed is a spring without mass or damping, the nonlinearity is given by Eq. (14), and the bore is an ideal cylinder with nearly lossless propagation. Figure 11 shows a three-level sister solution together with the related Helmholtz solution for a large number of harmonics, $N_p=2000$.

A solution of the lossless problem should satisfy the criteria (the dimensionless period being 2π) (Ref. 14).

$$\begin{cases} p(t+\pi) = -p(t) \\ u(t+\pi) = u(t), \end{cases} \tag{24}$$

which, for a square signal, is equivalent to the conditions stated before Eq. (15), noting that $p(t)=u(t)=0$ for all t is the static solution. It is easily verified graphically that both of the solutions in Fig. 11 satisfy these conditions. Moreover, since they also satisfy Eq. (14) at any time, the three-level solution is a solution of the lossless model.

Whereas the system of time-domain equations (24) has an infinity of solutions, truncation in the frequency domain limits the number of solutions. The unique solution of the HBM with only one harmonic is obviously a sine. Let us analyze the situation in the simplest nontrivial case of the lossless problem with two odd harmonics and a cubic expansion for nonlinear coupling. Ignoring even harmonics, the HBM gives a system of two equations (see Ref. 9)

$$\begin{cases} \alpha = 3P_1^2(1 + x + 2|x|^2) & \text{(25a)} \\ \alpha x = P_1^2(1 + 3x|x|^2 + 6x), & \text{(25b)} \end{cases}$$

where $\alpha=-A/C$ and $x=P_3/P_1$. As Eq. (25a) imposes P_3 to be real, solving this system amounts to solving

$$x^3 + x^2 - x = 1/3. \tag{26}$$

This equation has three real solutions: $x\simeq-1.5151$, -0.2776, and 0.7926. All of them are found by HARMBAL for negligible losses ($\eta=10^{-5}$), and the corresponding waveforms are presented in Fig. 12. We note that the second solution leads to the Helmholtz motion when increasing the number of harmonics (with the theoretical value known to be $x=-1/3$), whereas the third one corresponds to the three-level solution in Fig. 11. We can also easily imagine that these three solutions of the truncated problem are three-harmonic approximations of square waves that are distributed on three levels: $p^{\pm}\simeq\pm0.5$ and $p=0$. It should be noted that the conditions (24) for the continuous problem do not constrain the duration of each step. This has to be kept in mind when increasing N_p using the HBM.

While the Helmholtz motion is known to be stable,[14] the two three-level solutions can be considered as a combination of the static solution (the zero level) and the square wave

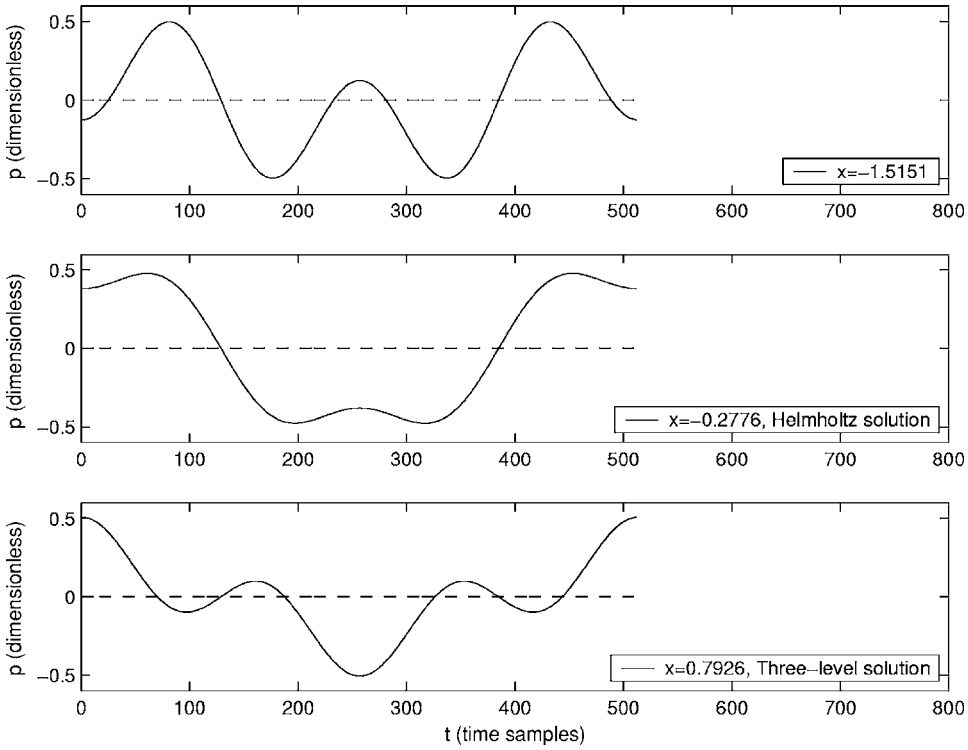


FIG. 12. The pressure waveform of the three solutions found by the HBM with $N_p=3$ employing the simple clarinet model with $\zeta=0.5$, $\gamma=0.4$, $\eta=10^{-5}$.

(two levels with opposite values). Since we know from Kergomard[14] that in the case of ideal propagation without losses the stability domains of these two solutions are mutually exclusive, it can be concluded that the three-level solutions are unstable.

Taking into account losses in the propagation does not make the three-level solutions vanish. But, a simple reasoning to determine the stability of this solution is not possible in this case. To the authors' knowledge, however, such a solution has never been observed experimentally at low level of excitation.

B. Initial value of the playing frequency

A practical difficulty encountered is the convergence of the playing frequency f_p. If its initial value is not close enough to the solution, divergence is almost inevitable. This occurs because the resonator impedance Z_r tends to vanish outside the immediate surroundings of the resonance peaks of the resonator, rendering $\mathbf{F}(\mathbf{P},f_p)$ very small and thereby $\mathbf{G}\simeq\mathbf{P}/P_1$ nearly constant with respect to f_p. The slope $\partial\mathbf{G}/\partial f_p$ thus becomes close to zero, the Newton step leads far away from the solution, and convergence fails. Dissipation widens the resonance peaks and thus also the convergence range.

For a simple system where the playing frequency is known to correspond to a resonance peak of the tube, initializing f_p is easy. However, with dispersion or other inharmonic effects, choosing an initial value for f_p may be difficult. In HARMBAL the problem may to some extent be avoided by the possibility of gradually adding the dispersion (or other inharmonic effects), so that the playing frequency can be followed quasicontinuously from a known solution without dispersion, for instance by using *hbmap*.

VI. CONCLUSIONS

The harmonic balance method (HBM) is suited for studies of self-sustained oscillations of musical instruments, and the computer program HARMBAL has been developed for this application. It is available with its source code,[11] has a free license, and is already in use by several researchers. It is programed in C, runs fast, and is easily used by other applications, such as for continuation purposes.

Some difficulties are related to the digital sampling of the signal and can be solved by introducing a backtracking mechanism. When using a large number of harmonics, the extreme case of the (lossless) Helmholtz motion can be solved for different shapes of resonators. Nevertheless, the value of the first harmonic P_1 seems to be well predicted by lower values of N_p, in particular close to the threshold of a direct bifurcation. For the saxophone we used a stepped-cone bore and observed one or more dips during the longest part of the period, depending on the number of steps. These dips approach pure impulses as N_p increases. The number of samples N_t in a period proved to be insignificant for these dips.

The HBM can lead to some alternative solutions for a unique set of parameters. The nondissipative versions of these solutions satisfy the continuous model equations, but they are not stable and thus cannot be attained by *ab initio* time-domain calculations. Another problem is the great sensitivity to the guessed playing frequency.

As a consequence, a certain expertise is needed in order to use the method, but, thanks to an automatic continuation procedure, the calculation is easy. We note that also experimental results can be used for the impedance of the resonator.

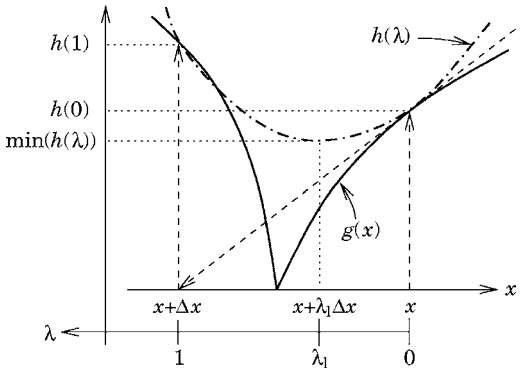


FIG. 13. Principle of backtracking in one dimension. The objective is to estimate the root of $g(x)$ (solid curve). Broken lines with arrows show how the Newton step Δx from x leads to divergence. $h(\lambda)$ (dot-dashed curve) is a second-order expansion of $g(x)$ along the Newton step, i.e., the λ axis. Minimum of $h(\lambda)$ should be closer to the root of $g(x)$ than $g(x+\Delta x)$.

ACKNOWLEDGMENTS

The European Union through the MOSART project is acknowledged for financial support. We would also like to thank Claudia Fritz at IRCAM in Paris for thorough testing and valuable feedback, Joël Gilbert at Laboratoire d'Acoustique de l'Université du Maine (LAUM) in Le Mans, and Philippe Guillemain at the Laboratoire de Mécanique et d'Acoustique at CNRS in Marseille for fruitful discussions during the work, and the latter also for kindly providing some MATLAB code for the time-domain model.

APPENDIX: BACKTRACKING SUMMARY

The principle of the backtracking algorithm[12] is illustrated in one dimension in Fig. 13, where $g(x)$ replaces $|\mathbf{G}(\mathbf{X},f)|$, although we use the multidimensional notation in the following. Defining the λ axis along the Newton step, we simply take a step $\lambda\Delta\mathbf{X}$ in this direction with λ between 0 and 1. The optimal value for λ is the one that minimizes the function $h(\lambda)$

$$h(\lambda) = \tfrac{1}{2}|\mathbf{G}(\mathbf{X}^i + \lambda\Delta\mathbf{X})|^2 \tag{A1}$$

with derivative

$$h'(\lambda) = (\mathbf{J}_G \cdot \mathbf{G})|_{\mathbf{X}^i+\lambda\Delta\mathbf{X}} \cdot \Delta\mathbf{X}. \tag{A2}$$

During the calculation of the failing Newton step, we computed $\mathbf{G}(\mathbf{X}^i)$ and $\mathbf{G}(\mathbf{X}^{i+1})$, so now it is possible to calculate with nearly no additional computational effort $h(0) = \frac{1}{2}|\mathbf{G}(\mathbf{X}^i)|^2$, $h'(0) = -|\mathbf{G}(\mathbf{X}^i)|^2$, and $h(1) = \frac{1}{2}|\mathbf{G}(\mathbf{X}^i+\Delta\mathbf{X})|^2 = \frac{1}{2}|\mathbf{G}(\mathbf{X}^{i+1})|^2$. This allows us to propose a quadratic approximation of h for λ between 0 and 1, for which the minimum is located at

$$\lambda_1 = -\frac{\frac{1}{2}h'(0)}{h(1) - h(0) - h'(0)}. \tag{A3}$$

It can be shown that λ_1 should not exceed 0.5, and in practice $\lambda_1 \geqslant 0.1$ is required to avoid too short a step at this stage.

If $|\mathbf{G}(\mathbf{X}^i+\lambda_1\Delta\mathbf{X})|$ still is larger than $|\mathbf{G}(\mathbf{X}^i)|$, $h(\lambda)$ is then modeled as a cubic function [using $h(\lambda_1)$ which has just been calculated]. The minimum of this cubic function gives a new value λ_2, again restricted to $0.1\lambda_1 < \lambda_2 < 0.5\lambda_1$. This calculation requires solving a system of two equations, so if λ_2 also is not accepted because $|\mathbf{G}(\mathbf{X}^i+\lambda_2\Delta\mathbf{X})|$ is still too large, we do not enhance to a fourth-order model of h, which would increase the computational cost much more. Instead, subsequent cubic modelings are performed using the two most recent values of λ. In practice, however, not many repetitions should be necessary before finding a better solution, if possible.

[1] H. L. F. Helmholtz, *On the Sensations of Tone* (Dover, New York, 1954), English translation of 4th German edition from 1877.
[2] M. E. McIntyre, R. T. Schumacher, and J. Woodhouse, "On the oscillations of musical instruments," J. Acoust. Soc. Am. **74**, 1325–1345 (1983).
[3] N. M. Krylov and N. N. Bogoliubov, *Introduction to Nonlinear Mechanics* (Princeton University Press, Princeton, NJ, 1947), English translation of Russian edition from 1936.
[4] M. Urabe, "Galerkin's procedure for nonlinear periodic systems," Arch. Ration. Mech. Anal. **20**, 120–152 (1965).
[5] A. Stokes, "On the approximation of nonlinear oscillations," J. Diff. Eqns. **12**, 535–558 (1972).
[6] M. S. Nakhla and J. Vlach, "A piecewise harmonic balance technique for determination of periodic response of nonlinear systems," IEEE Trans. Circuit Theory **23**(2), 85–91 (1976).
[7] R. T. Schumacher, "Self-sustained oscillation of the clarinet: An integral equation approach," Acustica **40**, 298–309 (1978).
[8] J. Gilbert, J. Kergomard, and E. Ngoya, "Calculation of the steady-state oscillations of a clarinet using the harmonic balance technique," J. Acoust. Soc. Am. **86**(1), 35–41 (1989).
[9] J. Kergomard, S. Ollivier, and J. Gilbert, "Calculation of the spectrum of self-sustained oscillators using a variable truncation method: Application to cylindrical reed instruments," Acta Acust. Acust. **86**(4), 685–703 (2000).
[10] C. Fritz, S. Farner, and J. Kergomard, "Some aspects of the harmonic balance method applied to the clarinet," Appl. Acoust. **65**(12), 1155–1180 (2004).
[11] S. Farner, HARMBAL, Computer program in C. Available on the Internet at http://www.lma.cnrs-mrs.fr/logiciels/harmbal/.
[12] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. (Cambridge University Press, United Kingdom, 1992).
[13] W. E. Worman, "Self-sustained nonlinear oscillations of medium amplitude in clarinet-like systems," Ph.D. thesis, Case Western Reserve University, 1971, Ann Arbor University Microfilms(ref. 71-22869).
[14] J. Kergomard, "Elementary considerations on reed-instruments oscillations," in *Mechanics of Musical Instruments*, Lecture notes CISM, edited by A. Hirschberg, J. Kergomard, and G. Weinreich (Springer, Vienna, 1995), Chap. 6, pp. 229–290.
[15] J.-P. Dalmont, J. Gilbert, and S. Ollivier, "Nonlinear characteristics of single-reed instruments: Quasistatic volume flow and reed opening measurements," J. Acoust. Soc. Am. **114**, 2253–2262 (2003).
[16] N. Grand, J. Gilbert, and F. Laloë, "Oscillation threshold of woodwind instruments," Acta Acust. Acust. **83**(1), 137–151 (1997).
[17] J.-P. Dalmont, J. Gilbert, and J. Kergomard, "Reed instruments, from small to large amplitude periodic oscillations and the helmholtz motion analogy," Acta Acust. Acust. **86**, 671–684 (2000).
[18] S. Ollivier, J. Kergomard, and J.-P. Dalmont, "Idealized models of reed woodwinds. II. On the stability of 'two-step' oscillations," Acta. Acust. Acust. **91**(1), 166–179 (2005).
[19] P. Guillemain, J. Kergomard, and T. Voinier, "Real-time synthesis models of wind instruments based on physical models," in Proc. of the Stockholm Music Acoustics Conference (SMAC) (Stockholm, Sweden, 2003), pp. 389–392.

Spectral characteristics of three styles of Croatian folk singing

Paul Boersma[a] and Gordana Kovacic
Institute of Phonetic Sciences, University of Amsterdam, Herengracht 338, 1016CG Amsterdam, The Netherlands

(Received 8 June 2005; revised 16 December 2005; accepted 31 December 2005)

This paper examines the differences between three Croatian folk singing styles, namely *klapa, ojkanje*, and *tarankanje*. In order to factor out singer-specific properties, each of the styles was performed by the same 12 professional male singers. The 36 performances were analyzed with a long-term average spectrum (LTAS) method from which direct effects of the pitch distribution were removed. After factoring out each singer's average, the 36 pitch-corrected LTAS contours were reduced to a two-dimensional representation in two ways: (1) a principal-component analysis and (2) a graphical plot of spectral slope versus speaker's formant strength. Both ways clearly separate the three styles. The spectrum of the klapa style turns out to be similar to that of speech. The ojkanje style is extremely loud and shows two spectral peaks: a sharp one tuned at twice the fundamental frequency and appropriate for long-distance communication on mountain slopes, and a broad one around 3.5 kHz, reminiscent of a speaker's formant. The tarankanje style has a very flat spectrum implemented by vocal pressedness and nasality, which is appropriate for blending into or imitating the timbral characteristics of the sopile folk instrument.
[DOI: 10.1121/1.2168549]

PACS number(s): 43.75.Rs, 43.70.Gr, 43.72.Ar [DD] Pages: 1805–1816

I. INTRODUCTION

Although there exist many more nonclassical than classical singers, most scientific information on the singing voice is based on studies of voices trained in the Western classical tradition, perhaps because its relative uniformity throughout the world allows a comparison of results across the various independent studies contributing to a deeper understanding of this single style. A broader understanding of the singing voice has to involve investigating the acoustics of many nonclassical styles, and some contributions have been made already on styles as diverse as pop (Schutte and Miller, 1993; Doskov *et al.*, 1995; Thalén and Sundberg, 2001; Borch and Sundberg, 2002), Broadway/musicals (Thalén and Sundberg, 2001; Stone *et al.*, 2003), country and western (Burns, 1986; Stone *et al.*, 1999; Sundberg *et al.*, 1999b; Cleveland *et al.*, 2001), jazz and blues (Thalén and Sundberg, 2001), Estonian folk (Ross, 1992), belting (Estill *et al.*, 1994), and overtone singing (Bloothooft *et al.*, 1992; Klingholz, 1993; Lindestad *et al.*, 2001; Van Tongeren, 2002).

The present investigation adds three more styles to the literature on acoustically investigated nonclassical singing: *klapa, ojkanje*, and *tarankanje*. These three very different styles all belong to the traditional music culture of Croatia. The relatively new klapa style (*klapa*="*group of friends*") originated in the 19th century in the Mediterranean part of Croatia (Dalmatia and Dalmatian islands), uses a Western European musical scale, and is usually performed *a cappella* with multiple parts in harmony, typically as soft, slow, serenadelike love songs (Rapanić, 1979; Ćaleta, 1997; Bezić, 1979). The tarankanje style is typical for the Istrian peninsula, the Kvarner islands, and the Croatian Littoral; it uses the Istrian musical scale, which has six narrowly spaced tones impossible to transcribe in the Western musical notation system (Bonifačić, 2001); it accompanies dance and is sung for a large part as strings of meaningless syllables (e.g., *tanana*) that can blend with or replace local wind instruments (Bonifačić, 1996). The ancient ojkanje style [the singers who participated in this study call it *dozivački*, from the verb *dozivati* "call (loudly)"], whose name refers to an *oj*-like syllable that is sung as a loud and tremolous "wild howl" before and/or after a loud short text (Dobronić, 1915; Bezić, 1968; Marošević, 1994), uses narrow non-Western intervals and tends to be perceived by outsiders as shouting or nonmusic (Dobronić, 1915; Ćaleta, 1999; Marošević, 2004); it is distinctive for mountainous Croatia, i.e., the Dinaric region and the Dalmatian hinterland.

The advantage of taking these three styles as the subject of investigation is that there exists a professional ensemble of folk singers that performs all of the three styles. Taking these singers as subjects for the present study allows us to reveal stylistic variation by analyzing the intrasubject differences across the performances of the three styles and factoring out any singer-specific characteristics.

II. METHOD

A. Subjects

A total of 12 male professional folk singers voluntarily took part in the investigation. All were members of LADO Folk Dance Ensemble of Croatia, which has been practicing song and dance from all regions of the country for over half a century. The singers had been performing Croatian folk music as LADO members for a period of 4 to 20 years, with an average of 10 years. Their ages ranged from 24 to 45, with an average of 33 years. None of the subjects had for-

[a]Electronic mail: paul.boersma@uva.nl

mally studied singing before joining the LADO ensemble. During the time of participation in the study, all singers reported to be in good vocal and physical condition (singer 1 reported just having recovered from a common cold, which is still audible in his klapa performance).

B. Data collection

The recordings were performed in an anechoic chamber of the Department of Electroacoustics of the Faculty of Electrical Engineering and Computing of the University of Zagreb. The background noise level measured inside the chamber was 19 dB(A) as measured by an integrating sound level meter (Brüel & Kjœr 2231).

The subjects were recorded one by one. Each was asked to perform one traditional song from each of the three styles. The songs were selected by the artistic director of the LADO Ensemble, who also chose the "key" for each song. The klapa style was represented by the song *Zaspalo je siroče* from Dalmatia, performed in G-major, the ojkanje style by the song *Mi smo rekli zapivati ode* from the Dalmatian hinterland, and the tarankanje style by the song *Homo u kolo* from Istria. Each singer performed each song three times, but only one performance of each song, namely the one that was judged best both by the singer himself and by the second author of the paper, was selected for acoustical analysis. The criteria were authenticity, stable vocal quality, and the singer's overall satisfaction with his performance.

Each singer performed in a standing position and was instructed to keep a constant distance of 0.3 m between his mouth and the microphone. The signal was recorded with a Behringer ECM 8000 omnidirectional microphone and fed via a TOA D-4 microphone preamplifier to an AIWA HD-S200 digital tape recorder with a sample rate of 44.1 kHz. Electroglottographic data were obtained in a fourth (shorter) performance with a Laryngograph, but these data turned out to be unreliable because of large vertical larynx movements (especially in ojkanje) and are not included in this study.

For each individual singer, the recording was preceded by a test recording in which the gain of the preamplifier was set to the optimal level for that singer. The gain was then kept constant for the three styles in order to make sure that the sound levels of the styles could afterwards be compared for each singer, although the recordings were not calibrated for absolute sound pressure level (for the unexpectedly loud ojkanje performance by singer 2, the recording gain was decreased by 6 dB, which was later corrected by doubling the amplitude).

The recording sessions thus yielded 12 performances of each of the three songs.[1] The average durations of the songs turned out to be 57.16 s for klapa, 50.75 s for ojkanje, and 43.50 s for tarankanje.

C. Acoustic analysis: Pitch-corrected LTAS

The purpose of the recordings was to obtain information on the average and individual production (phonatory and articulatory) and spectral properties (e.g., the presence or absence of a singer's formant) of the three styles. To this end, each of the 36 recordings was subjected to a long-term average spectrum (LTAS) analysis, performed with the phonetic analysis program PRAAT (Boersma and Weenink, 2005). Every LTAS was computed with a bin width of 125 Hz and a frequency range of 0–4 kHz. However, a simple LTAS may show undesirable F0-related phenomena. As an example consider Fig. 1, which shows the simple LTAS for tarankanje, pooled over all 12 singers.

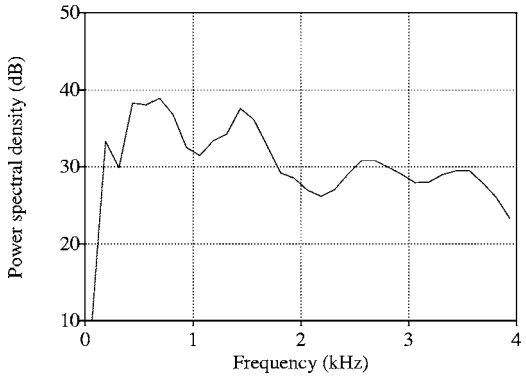


FIG. 1. Pooled LTAS for the 12 tarankanje performances, without pitch correction.

The figure clearly shows a peaked spectrum, especially in the second bin (125–250 Hz; the peak is at the center, i.e., at 187.5 Hz), the fourth bin (375–500 Hz), and the sixth bin (625–750 Hz). The individual singers vary in the presence of the second and third peaks, but the first peak is present in the LTASs of all 12 singers. The three peaks could correspond to the first, second, and third harmonics of a fundamental frequency around 220 Hz. This is confirmed by a histogram of the 70 544 F0 values measured by PRAAT for the 12 performances of tarankanje. Figure 2 shows that the largest peak is in the bin between 200 and 210 Hz.

To annihilate the influence of F0, a pitch-corrected LTAS method was designed, and this was used for most analyses in this paper.

The pitch-corrected LTAS procedure is summarized in Fig. 3 and runs as follows. For the voiced parts of the recording (Fig. 3 shows three examples), PRAAT's cycle-to-cycle waveform matching procedure detects all the pitch periods. Each pitch period is subsequently excised (data windowing is not needed, and the phase of the glottal pulse within the excised period does not influence the result). The figure shows the excision of one example period for each of the three voiced parts; the three periods have different dura-

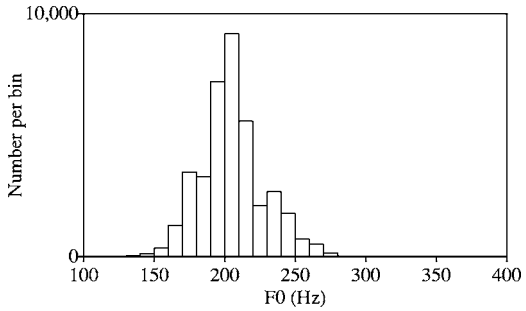


FIG. 2. F0 histogram for the 12 tarankanje performances.

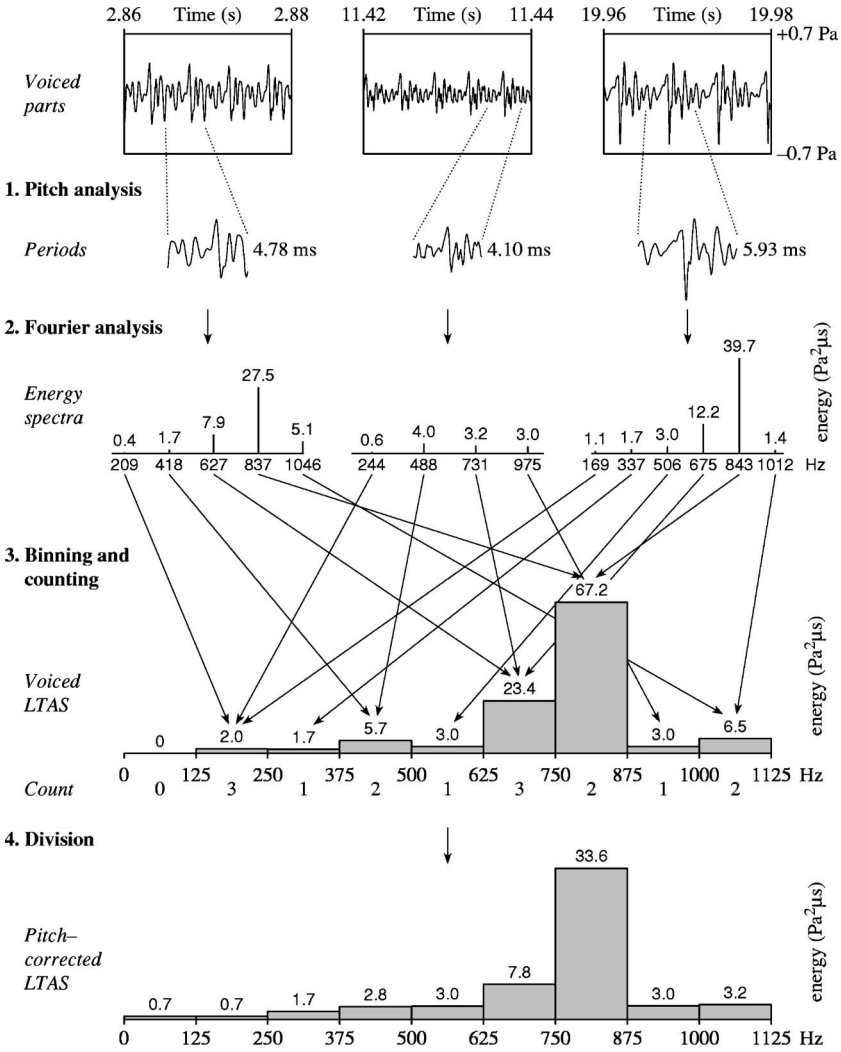


FIG. 3. The pitch-corrected LTAS method.

tions because the three voiced parts have different fundamental frequencies.

In the next step, each period is converted to a line spectrum by Fourier transformation. For instance, a period with a duration of 4.78 ms (one of the examples in Fig. 3) is transformed to a spectrum with a frequency spacing of 1/4.78 ms=209 Hz. This spectrum therefore contains information on energies at the harmonics, i.e., at 209, 418, 627, 837, and 1046 Hz and so on. For instance, the "energy"[2] of the third harmonic (at 627 Hz) is 7.9 $Pa^2\mu s$ (if the recording was indeed calibrated between −0.7 and +0.7 Pa, as the waveforms suggest).

The next step in the procedure is binning and counting. Each of the energies is put into the appropriate 125-Hz-wide bin of a LTAS. For the first period shown in the figure, the sixth bin, which runs from 625 to 750 Hz, receives 7.9 $Pa^2\mu s$ of energy, because the third harmonic at 627 Hz falls within this bin. Likewise, the high-energy peak of 27.5 $Pa^2\mu s$ associated with the fourth harmonic of 837 Hz is put into the seventh bin, which runs from 750 to 875 Hz. When the whole sound has been processed and all the energies have been added into their appropriate bins, the resulting histogram simply represents the LTAS of the voiced parts. However, while the energies in the harmonics of each period are put in their appropriate bins, the procedure also keeps track of how many harmonics leave their energies in each bin. For instance, after the processing of the three example periods there are three pieces of energy that have been put into bin 6 and two pieces of energy that have been put into bin 7. These counts can be read off Fig. 3 by counting the number of arrows that end in each bin and are also explicitly mentioned below the voiced LTAS histogram.

The last step is the actual pitch correction. The total energy in each bin is divided by the number of energies that had been put into that bin. The result is a LTAS in which each bin represents the average energy of the harmonics that entered it. For instance, the 23.4 $Pa^2\mu s$ that went into bin 6 is divided by 3 (the number of harmonics that contributed to the energy in this bin), yielding 7.8 $Pa^2\mu s$. The 67.2 $Pa^2\mu s$ that went into bin 7 is divided by only 2, because only two harmonics contributed to it. This enhances the peak in the seventh bin with respect to the value in the sixth bin.

To arrive at the final pitch-corrected LTAS, the procedure includes three more details. First, bins with a count of zero are subjected to linear interpolation between their neighbors or to constant extrapolation at the edge; in the figure this happens to the first bin. Next, in order to bring the histogram to the same scale as the uncorrected "voiced"

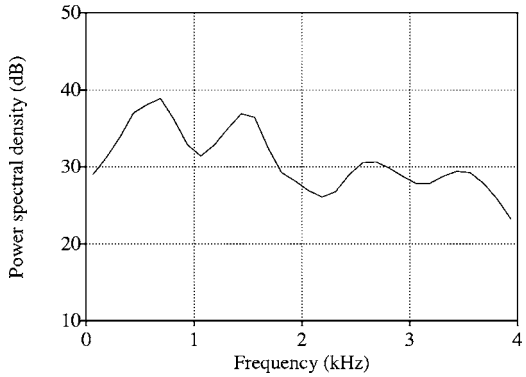


FIG. 4. Pooled LTAS for the 12 tarankanje performances, with pitch correction.

LTAS, the energy in every bin is multiplied by the average number of energies that went into a bin; in the example the multiplication factor would be 1.67, which is the total number of energies (15) divided by the number of bins (9). Finally, the energy in each bin is converted to an energy spectral density (in $Pa^2\,s/Hz$) by dividing it by the bin width (125 Hz), then converted to a power spectral density (in Pa^2/Hz) by dividing it by the total duration of the recording, then directly converted to a value in dB relative to 4.0×10^{-10} Pa^2/Hz. In this way, any calibration of the waveform in units of Pascal is faithfully preserved in the final pitch-corrected LTAS curve, and duration differences between the recordings are compensated for.

Figure 4 shows the final pitch-corrected LTAS curve for tarankanje. The curve is much smoother than the noncorrected LTAS of Fig. 1.

Comparable improvements apply to the other two styles, although the effect of the method is not always smoothing. Sharp spectral features can remain, as can be seen in the ojkanje curve in Fig. 5, which shows a deep and narrow valley around 800 Hz. To sum up, the pitch-corrected LTAS method maximally eradicates direct influences of F0 without sacrificing frequency selectivity. Some indirect influences of F0 (a rising F0 with a constant glottal waveform will lead to a stretching spectral envelope, and a high F0 may involve a raised larynx, which shortens the vocal tract and may thereby raise some formants) still remain visible in the pitch-corrected LTAS.

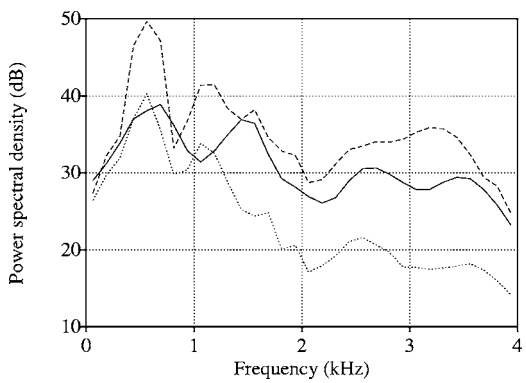


FIG. 5. Pooled pitch-corrected LTASs (—tarankanje,···klapa,---ojkanje).

III. RESULTS

A. Pooled data

Figure 5 shows the pitch-corrected LTASs for each of the three styles, pooled over all 12 singers. Several interstyle differences can be read off this figure. The most noticeable differences are related to sound levels, spectral slopes, and locations of the spectral peaks.

As far as the sound levels are concerned, Fig. 5 shows that klapa is least loud and ojkanje is loudest. This feature is noticeable through the differences between the amplitudes of the strongest spectral peaks at around 0.6 kHz where both the klapa and tarankanje styles have an approximately 10 dB lower amplitude than the ojkanje style. This corroborates ethnomusicological descriptions of the ojkanje style that state that vocal loudness is a dominant feature of this style (Ćaleta, 1999).

The second difference among the styles is in the average slope of the spectrum. The spectral slope is related to the relative speed of glottal closure (Fant, 1960, p. 270), which again is correlated to the vocal loudness mentioned in the previous paragraph [in fact, spectral slope is a main auditory cue for perceived loudness in speech (Sluijter, 1995)]. The spectral slope is indeed steepest for klapa, which also has the lowest vocal intensity; the difference of 20 dB between the low- and high-frequency regions resembles what is generally found in speech (Kuwabara and Ohgushi, 1984; Leino, 1994; Cleveland *et al.*, 2001). Likewise, the spectral slope is flatter for ojkanje, which has the largest vocal intensity; the low-high difference of 15 dB corresponds to that in very loud speech or in shouting (Nawka *et al.*, 1997; Nordenberg and Sundberg, 2003). The tarankanje style is special: it has an unusually flat slope (10 dB) but medium to loud vocal intensity. The flat slope suggests that this style employs a pressed voice (Stevens, 1998, p. 85; Bergan *et al.*, 2004, p. 311).

The third difference between the three styles is in the location and regularity of spectral peaks. The klapa and ojkanje styles have peaks around 0.6 and 1.1 kHz, which correspond to the locations found in speech (of any vocal intensity, including shouting). The tarankanje style is again very different: the location of the two peaks at 0.7 and 1.5 kHz, together with the valley at 1.1 kHz, indicates that the average spectrum corresponds to that of the nasalized low front vowel [æ̃] (Stevens, 1998, p. 311), while the strength of the second peak (almost as high as the first) is probably due both to the pressed voice quality mentioned before and to the raising of the bandwidth of the first formant as a result of nasality (Stevens, 1998, pp. 310, 312).

For the klapa and the tarankanje styles there seems to be a regular pattern of peaks appearing clearly around 2.5 and 3.5 kHz, probably reflecting the third and fourth formants (F3 and F4). This is the spectral region where one could look for the *singer's formant*, a strong resonance phenomenon at about 2.8 kHz, typical of operatic singing voices (Bartholomew, 1934; Sundberg, 1973, 1974), but none of the three spectra show such a strong peak in this area. However, the ojkanje style is characterized by a prominent broad plateau between 2.2 and 3.8 kHz whose local peak is suggestive of a *speaker's formant* (or *actor's formant*), a resonance phenom-

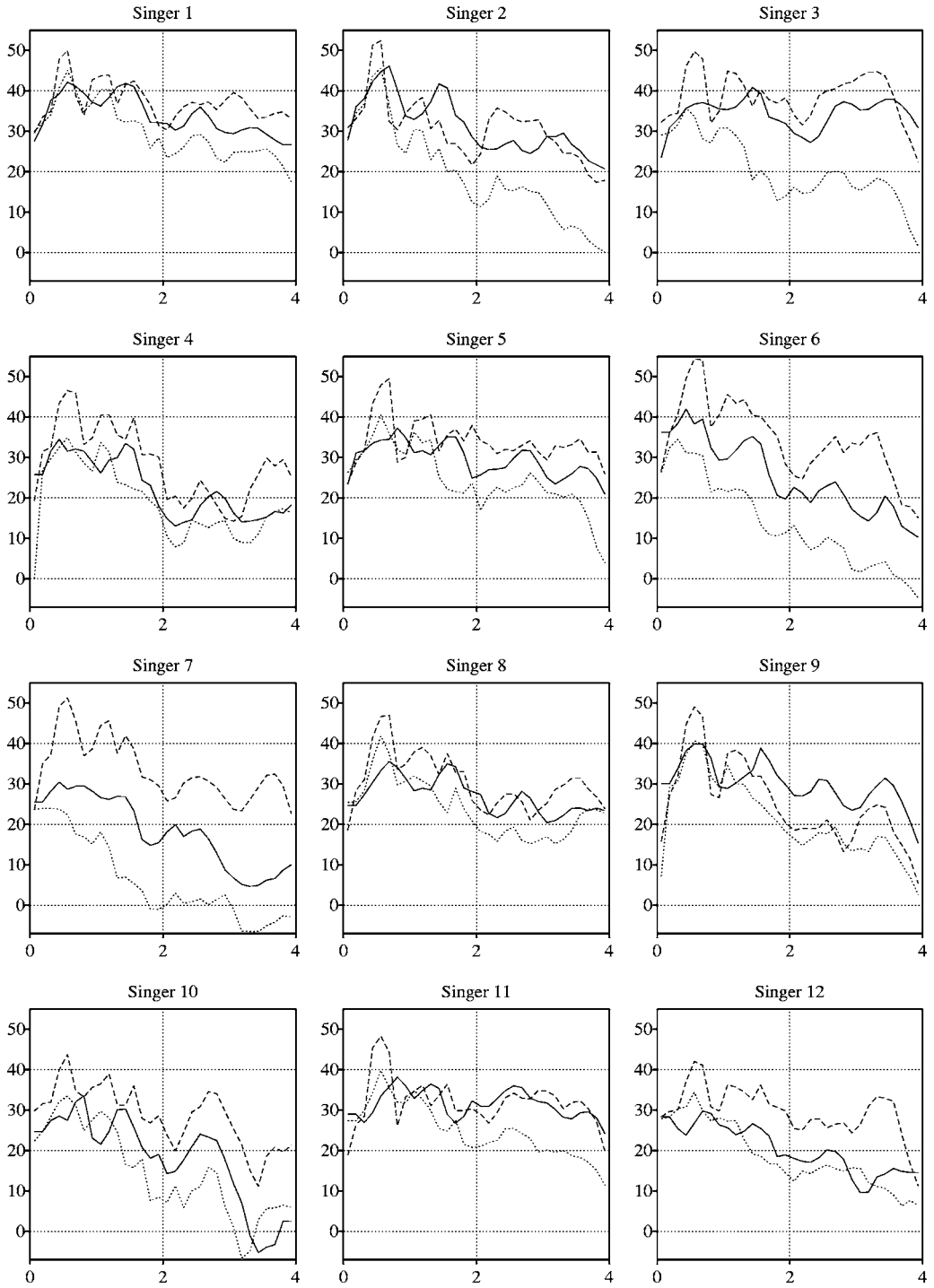


FIG. 6. Individual pitch-corrected LTASs (—tarankanje,···klapa,---ojkanje). Horizontally: frequency in kHz; vertically: power spectral density in dB.

enon that is usually associated with the speech of professional speakers such as radio announcers (Kuwabara and Ohgushi, 1984) and theater actors (Leino, 1994; Nawka *et al.*, 1997; Thunberg, 2003). These authors agree that a speaker's formant is weaker than a singer's formant and associated with a higher spectral region than the singer's formant, namely between 3.0 and 3.8 kHz (for males). In the present study, the amplitude of the peak in ojkanje is about 17 dB lower than the strongest peak of the spectrum. The naming of this peak as a speaker's formant may be problematic, though, since it is found in singing rather than in speaking. Section IV proposes an alternative term.

B. Individual data

While the average interstyle differences can be read off Fig. 5, it is also important to investigate to what extent the differences are consistent across speakers. This can be done informally by visual inspection of the individual pitch-corrected LTASs, and formally by performing a principal-component analysis or a data reduction into preestablished properties of the individual LTASs.

Figure 6 shows the individual pitch-corrected LTASs for each of the 12 singers and each of the three styles. Several properties that were noted for the pooled data of Fig. 5 also

appear for many, most, or all of the individual singers: the lowest vocal intensity for klapa (all singers except perhaps singers 4 and 8), the highest vocal intensity for ojkanje (all singers except perhaps 9), the steepest (i.e., most speechlike) slope for klapa (all singers except perhaps 9), the flattest slope for tarankanje (singers 2, 3, 5, 8, 9, and 11), the two speechlike peaks around 0.6 and 1.1 kHz for klapa and ojkanje (all singers), the valley around 1.1 kHz for tarankanje (all singers except 7 and 12), and the speaker's-formant-like broad spectral peak around 3.5 kHz for ojkanje (singers 3, 4, 6, 7, 8, 9, 12, and perhaps 1 and 11). Apart from these main effects of style, the figure shows main effects of speaker (e.g., singer 6 has steeper spectral slopes than singer 1, for all three styles, and singer 10 uses a singer's-formant-like local peak in all styles) and interactions between speaker and style (e.g., singers 6 and 7 make larger loudness differences between the styles than singers 1, 8, and 11).

The main interest here must be in the detectability of the styles on the basis of the differences noted in Sec. III A, i.e., in the consistency between singers with regard to these differences. Two methods are discussed in Secs. III C and III D.

C. Style discrimination by principal component analysis

While the consistency across singers can be informally read off the individual data, there is also a formal technique that can establish this consistency. Intersubject consistency in the relation between the styles is present if, after any singer-specific properties have been factored out, the remaining variation between the 36 recordings is mainly due to the style. Precisely this can be measured by a principal component analysis. The first step is to factor out singer dependency by computing for each singer his average LTAS (i.e., averaged over the three styles) and subtracting this average from each of his three LTASs. The resulting singer-normalized LTASs of the three styles thus add up to zero for each speaker. The 36 singer-normalized LTASs can be regarded as 36 vectors in 32 dimensions (32 is the number of 125-Hz bins in the 0–4 kHz range). The first two principal components of these 36 vectors are shown in Fig. 7.

The first principal component is a slightly rising line, lying entirely above zero. This means that a performance that contains this first component to a high degree is one that combines an overall high intensity level with relatively strong high-frequency components. Since these two spectral features can both be related to loud singing, the fact that the first principal component has this particular shape shows that most of the variation between the 36 singer-normalized LTASs (in fact, 81.2%) is variation in the loudness of the voice. The second principal component (accounting for 6.3% of the variation) has positive peaks around 550 and 1100 Hz and negative peaks around 850 and 1450 Hz. The two negative peaks correspond to the regions in Fig. 5 where the tarankanje curve is higher then the ojkanje curve, and the positive peaks correspond to the regions where the tarankanje curve lies deepest below the two others. The second component, then, turns out to measure the degree to which a performance contains the tarankanje-specific coloring of the spectrum below 2 kHz.

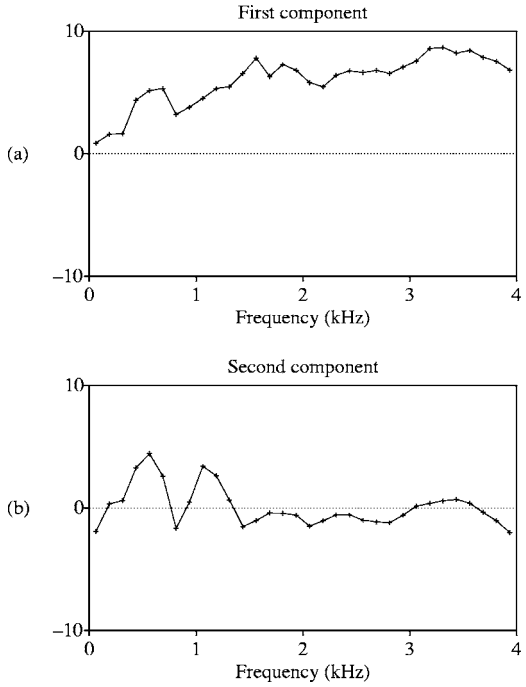


FIG. 7. The first two principal components of the 36 pitch- and speaker-corrected LTASs.

The interesting thing now is to see where the 36 singer-normalized LTASs end up in the two-dimensional space spanned by these two principal components. The result is in Fig. 8. Each of the 12 singers occurs three times in this figure. The three marks labelled "10," for instance, represent the tarankanje, klapa, and ojkanje performances of singer 10; as a result of the singer correction, the horizontal and vertical averages of these three points are zero.

The klapa performances tend to have negative values for the first component, which was expected because Fig. 5 shows that the average klapa performance is less loud [i.e., less like Fig. 7(a)] than the average tarankanje or ojkanje performance. The tarankanje performances tend to have

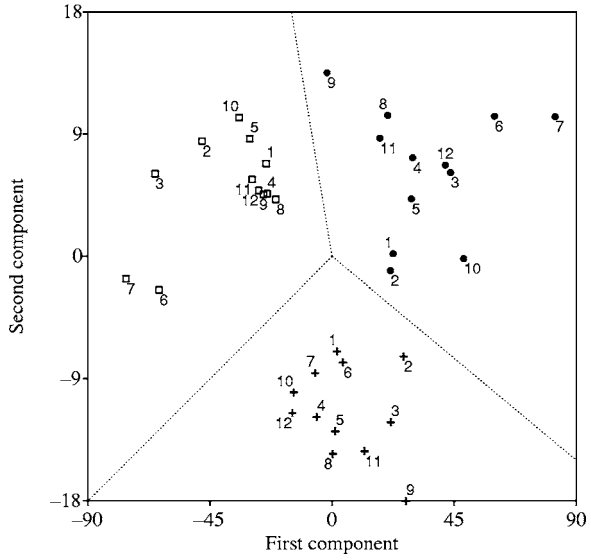


FIG. 8. Principal component analysis of the 36 performances. Pluses =tarankanje, rectangles=klapa, dots=ojkanje, 1,...,12=the 12 singers.

negative values for the second component, which was expected because the negative peaks in Fig. 7(b) correspond to positive peaks in Fig. 5 for tarankanje.

The main finding of Fig. 8 is that the styles form clusters without overlap: although the principal component analysis is not a clustering algorithm and does not know what style was associated with what LTAS curve, it turns out to be able to draw a perfect linear division among the 36 performances. It must be noted that, without singer normalization, this separability would not have occurred: if the 36 recordings are regarded as stemming from 36 different singers, the three clouds in Fig. 8 will come to overlap considerably. Thus, the separability of the three styles in the present PCA experiment is a direct result of the experimenters' decision to investigate the three styles with the same 12 singers.

D. Style discrimination by spectral slope and speaker's formant

Whereas the principal component analysis was allowed to take into account the overall intensity level of each performance and to define its own dimensions, the 36 performances can also be plotted in a predefined space of two dimensions that do not refer to acoustic intensity level. It seems reasonable to choose one dimension that reflects a characteristic of the voice source and one dimension that reflects a characteristic of supralaryngeal articulation (i.e., of the shape of the vocal tract).

For the voice source dimension it was decided to measure the *global spectral slope*. Spectral slope (or *spectral tilt*) measures are related to several characteristics of glottal performance, such as vocal intensity (Glave and Rietveld, 1975; Gauffin and Sundberg, 1989; Kiukaanniemi *et al.*, 1982; Sundberg, 2001, pp. 177–178) and hypo- and hyperfunctionality (Löfqvist and Mandersson, 1987). The global slope measure was defined as the difference of the average sound pressure level below 1.0 kHz and the average level between 1.0 and 4.0 kHz. The simple method of determining the spectral slope by using a pivot of 1 kHz is due to Frøkjær-Jensen and Prytz (1976), who directly computed the difference in energy above and below. However, a measure based on perceptual loudness is likely to reflect the psychoacoustic spectral slope better than an energy measure would (Zwicker and Feldtkeller, 1967) and is therefore more likely to reflect style discrimination by humans. The sound pressure levels, then, are computed in dB but mediated by sone units. The global spectral slope is then

$$10\log_2\frac{1}{3.0\text{kHz}}\int_{1.0\text{kHz}}^{4.0\text{kHz}}2^{\text{PSD}(f)/10}\,df - 10\log_2\frac{1}{1.0\text{kHz}}\int_{0}^{1.0\text{kHz}}2^{\text{PSD}(f)/10}\,df, \quad (1)$$

where PSD(f) is the power spectral density (in dB) at frequency f, as estimated from the height of the corresponding bin of the pitch-corrected LTAS.

For the articulatory dimension it was decided to measure the strength of the "speaker's formant," i.e., the strength of the spectral region between 3.0 and 3.8 kHz that "good" (cultivated, trained) speakers have been found to enhance (see Sec. III B). The speaker's formant is thought to have an articulatory correlate (Nolan, 1983, p. 151), perhaps a clustering of F4 and F5 (Leino, 1994), or just F4 alone, supported by a long closure phase (Cleveland *et al.*, 2001). One could define a global speaker's formant strength by subtracting the height of the 3.5-kHz peak from the height of a peak below 1 kHz [e.g., Nawka *et al.*, 1997; cf. the similar "resonance balance" method for the singer's formant by Schutte and Miller (1984)], but as Leino (1994, p. 209) points out, such a definition would confound this articulatory measure with the global spectral slope, which is related to the voice source (in the present case, such a subtraction would lead to low values for both tarankanje and ojkanje, but by different causes). In order for the measure of speaker's formant strength to be as independent of the spectral slope as possible, it has to be taken relative to the level of the surrounding spectral region, i.e., the regions between 2.2 and 3.0 kHz and between 3.8 and 4.6 kHz. The formula for this *local speaker's formant strength* is analogous to the formula for the global spectral slope:

$$10\log_2\frac{1}{0.8\text{kHz}}\int_{3.0\text{kHz}}^{3.8\text{kHz}}2^{\text{PSD}(f)/10}\,df - 10\log_2\frac{1}{1.6\text{kHz}}\left(\int_{2.2\text{kHz}}^{3.0\text{kHz}}2^{\text{PSD}(f)/10}\,df + \int_{3.8\text{kHz}}^{4.6\text{kHz}}2^{\text{PSD}(f)/10}\,df\right). \quad (2)$$

After the computation of the global spectral slope and the local speaker's formant strength, the values of these two numbers are normalized for each singer, analogously to the singer normalization performed before in the principal component analysis. Figure 9 plots all 36 performances in the space spanned by the two singer-normalized dimensions just defined.

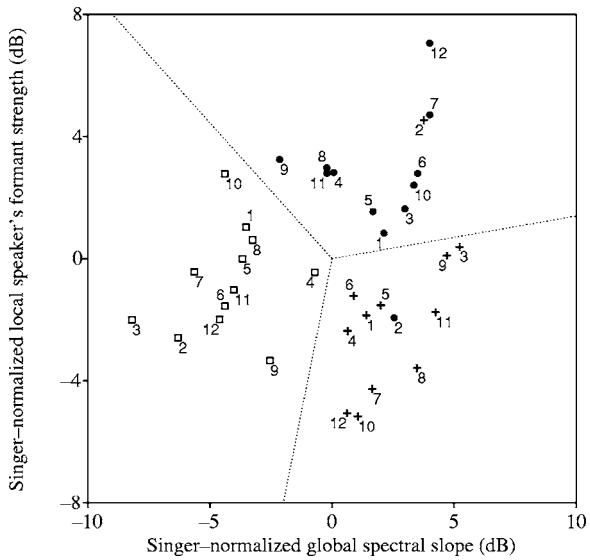


FIG. 9. The locations of the 36 performances in a space of two predefined spectral shape characteristics. Pluses=tarankanje, rectangles=klapa, dots =ojkanje, 1, … , 12=the 12 singers.

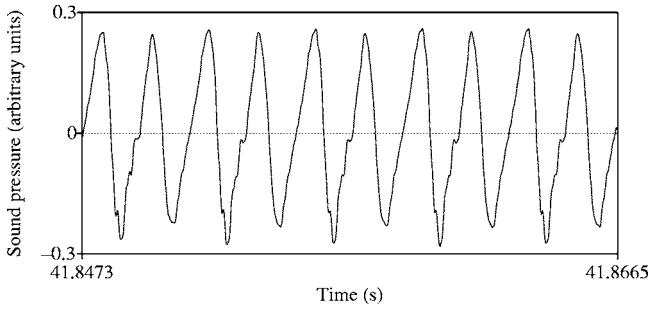


FIG. 10. Five periods of a vowel in ojkanje (singer 9).

Figure 9 shows that several differences between the styles are consistent across singers: the fact that every singer's klapa performance turns up at the left of both his tarankanje and his ojkanje performance means that every singer's spectral slope for klapa is greater than his spectral slope for either other style. The fact that all tarankanje performances show up at the right of the zero vertical shows that every singer's tarankanje performance has a flatter slope than his average performance has. The fact that 11 ojkanje performances show up above the zero horizontal shows that for every singer except singer 2 the speaker's formant for ojkanje is stronger than his average speaker's formant. Likewise, the speaker's formant seems to be absent from the tarankanje performances of all singers except singer 2.

The overall image of Fig. 9 is that a linear separation between the three styles is very well possible, i.e., given three performances of the same singer the singer-normalized measurements of spectral slope and speaker's formant are capable of predicting which performance belongs to which style (except for singer 2). This separation is shown by the dividing lines in Fig. 9. The two dimensions are probably independently controllable by the singer: the spectral slope is controlled by the voice source, whereas the speaker's formant is controlled by supralaryngeal articulation.

As in Fig. 8, the separability of the three styles in Fig. 9 is a direct consequence of investigating a single group of 12 singers on different styles, because without singer normalization the three clouds would overlap to a large degree.

E. Style discrimination by the strength of the second harmonic

There exist spectral characteristics that cannot be read off directly from the LTAS curves. One of them relates to the rather narrow peak around 600 Hz in the ojkanje style. Since the fundamental frequency of the ojkanje song investigated here is around 300 Hz, it may be worthwhile to investigate the hypothesis that ojkanje singers aim at maximizing the amplitude of the second harmonic. Figure 10 shows five periods of the ojkanje performance by singer 9. The F0 is 262 Hz, but the waveform resembles a 524 sine wave, so that it is likely that most of the spectral energy is in the second harmonic. Preliminary investigations into authentic ojkanje singers suggest that the singers aim at lending prominence not to a specific frequency, but to the second harmonic.

To investigate how well LADO singers succeed in bringing forward the second harmonic (H2), it was computed how high the energy of the second harmonic rose above the average energy of the first harmonic (i.e., F0) and the third. The measurement method can be exemplified with Fig. 3: to compute the total energy associated with H2, for instance, one just adds the energies associated with the second harmonics of the three periods in the figure, yielding $1.7+4.0+1.7=7.4$ Pa^2 μs. For the 12 LADO singers, then, the local H2 strength [i.e., $H2-(H1+H3)/2$] turned out to lie between 10.2 and 19.4 dB. These high values positively identify ojkanje, even without singer normalization, since the two other styles have lower H2 strengths (tarankanje: between −4.2 and +4.5 dB; klapa: between 0.1 and 10.1 dB). With singer normalization, the discrimination improves: every singer's H2 strength for ojkanje is much higher than his H2 strength for klapa; the difference lies between 5.3 and 14.4 dB, with a median difference of 9.2 dB.

Another way of establishing the strength of the second harmonic starts with realizing that a strong second harmonic of a male singer with an F0 of 220 to 320 Hz is likely to be the result of a tuning of the first formant. Figure 11 shows the distributions of the first four formants for each of the three styles, pooled over all 12 singers (the formants were determined with the PRAAT program, requiring five formants be-

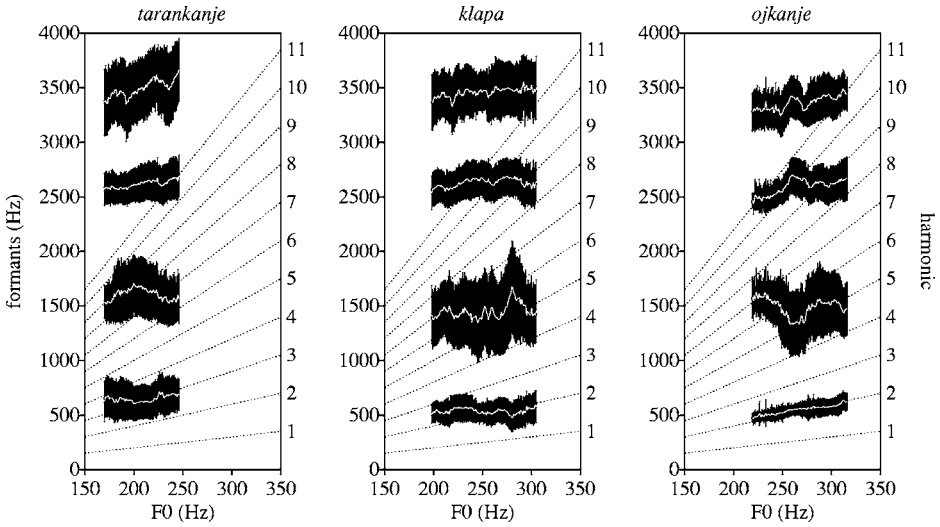


FIG. 11. Pooled formant distributions as functions of F0.

low 5000 Hz, with an analysis window length of 25 ms). The figure contains information only on those moments in the performance where F0 is between the 5th and 95th percentile of the pooled F0 distribution of the style and all four formants are simultaneously between the 5th and 95th percentiles of their pooled distributions. The white curves, then, indicate the average formants as functions of F0; these curves were computed by averaging approximately 30 000 formant values (per style per formant) in 1-Hz-wide bins. The black areas above and below the curves (they are actually vertical lines, one line per Hz) indicate the standard deviations of the formants as functions of F0. The dotted lines in the figure indicate the first to eleventh harmonics as functions of F0.

The figure confirms the hypothesis. The F1 curve for ojkanje closely follows the line of the second harmonic. This cannot be a side effect of the analysis method, because the other styles show hardly any dependence of F1 on F0. The standard deviation of F1 for ojkanje is seen to be very small. This, together with the coincidence of F1 with H2, proves that these singers tune their F1 to 2F0 (or the reverse, i.e., 2F0 to vowel height). This ojkanje song can therefore be seen as a kind of overtone singing.

A third thing that would contribute to the sharpness of the peak is the bandwidth of the first formant. The LADO singers did not utilize this possibility in this song: the median F1 bandwidth was 179 Hz, which was not so different from that of the other two styles (tarankanje 183 Hz, klapa 130 Hz).[3]

IV. DISCUSSION

In this section we want to point out that the differences between the styles, as observed in Sec. III, have perceptual goals: the acoustic differences do not just reflect single articulatory differences. Rather, for each style there are a multitude of articulatory features that synergetically act to achieve a specific psychoacoustic effect. The following subsections explain this in detail for each style.

A. Tarankanje: Whiteness

The extraordinary flatness of the spectrum in the tarankanje style seems to be implemented synergetically by settings of the voice source and by supralaryngeal settings. The energy difference between the 0–2 kHz and 2–4 kHz regions is made as small as possible by producing a pressed voice. The energy difference between the first and second spectral peaks (at 0.7 and 1.5 kHz) is made as small as possible by maintaining a nasal vocal quality reminiscent of an open nasal front vowel (Stevens, 1998, pp. 310–312). The song under investigation here seems to aim at enhancing this effect by in fact using the nasalized low front vowel [æ̃] as its most frequent vowel, although the spectrum turns out to stay very similar if all [næ̃]-like syllables are removed (which shows that the phonemic makeup of the song is largely irrelevant for the spectrum). Indeed, tarankanje singing has been described as "strong and partially through the nose, so the tone colour is nasal" (Bonifačić, 2001, p. 75).

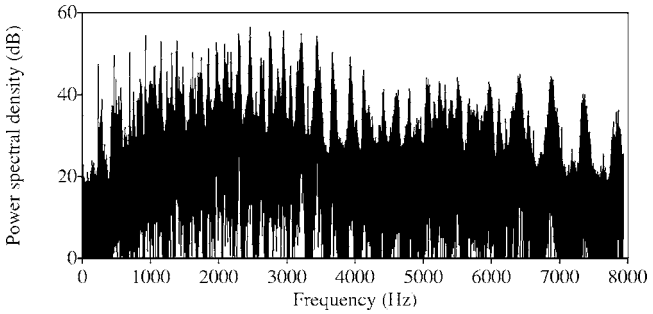


FIG. 12. The flat spectrum of two sopile instruments.

Together, these production tricks implement a spectrum that is as flat as possible, i.e., what Bergan *et al.* (2004) call a relatively "white" sound. This flat spectrum renders the sound similar to that of the wind instruments that are part of folk tradition, namely *sopile* (shawms) or *roženice*, which have been described as producing a piercing and nasal tone quality (Bonifačić, 2001). An example of a spectrum of a large and a small sopile playing together is given in Fig. 12. According to Bonifačić (1996), imitating these instruments is the very goal of tarankanje singing: if the instrument is not available, tarankanje singing can still take over its role as an accompaniment to dance.

B. Ojkanje: First-formant tuning and a speaker's (shouter's) formant

The ojkanje style is characterized by great loudness, which is reflected as a generally higher level of intensity, by the tuning of the second harmonic to F1, and by a broad spectral peak for high frequencies, reminiscent of a speaker's formant.

The tuning of the first formant as the first overtone of the fundamental frequency in ojkanje is by far the strongest peak in the spectra of all styles together. Ćaleta (1999) describes ojkanje as shoutinglike vocal production in a high register. The label of "shouting" is common and refers both to the listener's impression of this style and to the original perceptual goal of this style, which has been described as the need to communicate over a great physical distance in sparsely settled mountainous areas (Dobronić, 1915; Bezić, 1968; Marošević, 2004). The frequency range from 600 to 800 Hz is indeed optimal for achieving the special carrying power of ojkanje, because for lower frequencies the human ear becomes less sensitive (Fletcher and Munson, 1933) and for higher frequencies air absorption may hamper transmission (Evans *et al.* 1971). A very similar sharp peak has been found before in *kölning*, which is a Swedish singing style (the name is reported to derive from the verb *kalla* "call") that maids used for calling cattle and to communicate to each other, in mountainous areas (Johnson *et al.* 1985; Sundberg, 1995, p. 2). In kölning women tune their F1 to their F0 (Johnson *et al.*, 1985, p. 198). With a typical difference of an octave between women's and men's singing, it seems only natural that men would tune their F1 to 2F0, women to F0. Generalizing from the two cases of ojkanje and kölning, and following the tradition in voice research to call spectral

peaks "formants," one could call this tuned F1 peak the "low shouter's formant."

The second, broader, peak lies between 3.0 and 3.8 kHz and is probably identical to the speaker's formant. Besides in "good" speakers, such a peak has been attested before in the speech and song of country singers (Cleveland *et al.*, 2001). It can be interpreted as the combined result of a vocal action and an articulatory posture. The vocal action is the observed high vocal intensity, a thing that tends to affect the level of high-frequency peaks more than the level of lower frequencies; this correlation has been shown in singing (Bartholomew, 1934; Hollien, 1983; Bloothooft and Plomp, 1986; Sundberg, 2001) as well as in speech (Nawka *et al.*, 1997; Nordenberg and Sundberg, 2003) and is present in the comparison of ojkanje with the speechlike klapa style as well (Fig. 5). The articulatory posture must be the posture that leads to a strong or close F4 and F5, hence a strong local peak (see Sec. III D). In ojkanje, the loudness in general and the high-frequency strength in particular are probably implemented synergetically by various respiratory, phonatory, and articulatory settings, namely a high subglottal pressure, a long closed phase of the vocal folds, a raised larynx, and a wide opening of the jaw [cf. Johnson *et al.* (1985), who found that kölning comes with a high subglottal pressure, a raised larynx, and a large jaw opening].[4] If future research attests the high-frequency peak in similar calling styles around the world, the term *shouter's ring* or *shouter's formant* might become appropriate.

The aforementioned proposal that the perceptual goal of ojkanje is the need to communicate over a large distance in mountainous areas (Dobronić, 1915; Bezić, 1968; Marošević, 2004) has now been confirmed by the observation of a sharply tuned resonance ($F1=2F0_{male}$) in the farthest-carrying frequency range. The proposal is also corroborated by Marošević's (2004) observation that similar vocal styles are found in mountainous regions elsewhere (Albania, Bulgaria, Greece, Turkey). And the proposal is especially strongly supported by Johnson *et al.*'s (1985) proposal that the same perceptual goal applies to a calling song (kölning) with a very similar sharply tuned resonance ($F1=F0_{female}$).

C. Klapa: Harmony

The klapa style has been described as sounding pleasant and beautiful, even cultivated (Bezić, 1979). Its perceptual goal has been described as "to achieve the best possible blend of chords" (Ćaleta, 1997, p. 135), probably coinciding with that of other love songs all over the world. Klapa singers use two tricks to achieve harmony (in the sense of overtone matching): first, they sing in a harmonically based (Western-like) musical scale; second, they use multiple larynges, i.e., klapa is always performed in a choir consisting of five to eight singers.

D. The absence of a perceptual goal: Singer's formant

The current study showed that none of the three singing styles relies on a singer's formant.[5] This finding can be compared to the findings by Burns (1986) for American country-and-western folk singing, by Ross (1992) for Estonian folk singing, and by Cleveland *et al.* (2001) for country singing, none of whom found any evidence for a singer's formant in folk singing styles. This absence is understandable in the light of Sundberg's (1972) proposal that the singer's formant originally developed in order for the singer not to be masked by the spectrum of a symphonic orchestra.[6] According to Sundberg, singers would have no need for a singer's formant if they are accompanied by instruments with lower sound levels, such as a lute. This can explain why a singer's formant never developed in ojkanje, which is always performed *a cappella*, nor in klapa, which is usually sung *a cappella* and only rarely with the accompaniment of light string instruments such as mandolins. The tarankanje style is typically performed with two complementary voices [so-called "big" and "small" (Karabaić, 1956)], which may be two human voices or one human voice and a sopile instrument. In the latter case, the human voice is not meant to overcome the spectrum of the sopile. On the contrary, the perceptual goal of the human voice seems to be "onomatopoeic imitation" (Bonifačić, 1996) to mimic the spectral whiteness of the sopile instrument.

In the case of klapa, the absence of a singer's formant can partly be explained by the fact that it is performed in a choir: since multiple voices contribute to the loudness, the singer's formant is superfluous. Rossing *et al.* (1986) indeed showed that professional singers use a singer's formant consistingly mainly when performing solo, and use it less when performing in a choir.

V. CONCLUSION

The goal of the current study, namely to find spectral differences between three styles of Croatian folk singing, was successfully reached because it turned out that spectral differences between the styles could be established for each of the 12 singers in the same way (except for singer 2's tarankanje and ojkanje in Fig. 9).

The spectral characteristics of the three styles were ultimately explained in detail by their original or current perceptual goals. The absence of the perceptual goal of overcoming an orchestra explained the lack of a singer's formant in all three styles, either negatively (because of the lack of an orchestra) or positively (because of the very desire to blend with the instrumental accompaniment). The presence of the perceptual goal of being heard across large distances in mountainous areas explained the F1 tuning in ojkanje. Finally, the presence of the perceptual goal of imitating the sound of the sopile explained the flat ("white") spectrum in tarankanje.

For all three styles the present paper investigated the same 12 singers. The objective of this was to be able to perform singer normalization, and indeed this normalization turned out to be a requirement for the near-perfect separability of the three styles observed in Figs. 8 and 9 for the professional singers of the LADO ensemble. Doing the tests of the present paper with performers randomly sampled from three hypothetical populations of singers with LADO-like interpretations of the styles is therefore expected not to yield

such a good separability. It is feasible, however, to redo the tests with authentic singers with specialized skills in a single style, who may compensate for the lack of normalizability by having more outspoken style features. Informal observations indeed suggest that authentic ojkanje performers shout even higher and louder while at the same time expending less effort, authentic tarankanje performers match the sound of the sopile even more closely, and authentic klapa singers sing even more harmonically and smoothly.

ACKNOWLEDGMENTS

The research was supported by a Huygens scholarship 2002/03 awarded by Nuffic. The authors are thankful to LADO Folk Dance Ensemble of Croatia for the willingness and patience to participate in the study despite a very tight schedule at the time of data collection, and also to the staff of the Department of Electroacoustics of the Faculty of Electrical Engineering and Computing of the University of Zagreb (Croatia) for their assistance throughout the above-mentioned phase of the investigation.

[1]See EPAPS Document No. E-JASMAN-119–059603 for the sound files of the 36 performances. This document can be reached through a direct link in the online article's HTML reference section or via the EPAPS homepage (http://www.aip.org/pubservs/epaps.html).

[2]The unit is Pascal-squared times microseconds. To obtain an average sound intensity in W/m^2, one divides by the duration of the period and by the acoustic impedance ρc, where ρ is the density of air and c is the speed of sound.

[3]A performance of the song *Oj djevojko* by authentic singers turned out to have an extremely low F1 bandwidth of only 10 Hz, leading to an H1-H2 difference of −30 dB in the acoustic signal. To what extent the difference between 179 and 10 Hz is due to a difference between the songs, the regions, and/or the performers must be left for future research.

[4]A direct correlation between subglottal pressure and duration of the closure phase has been reported for untrained speakers (Sundberg *et al.*, 2005) and for professional singers (Sundberg *et al.*, 1999a). However, it is possible that the correlation appears only if the goal of the performer is to vary loudness, as it is in the case of both ojkanje and the participants in Sundberg *et al.* (1999a) and Sundberg *et al.* (2005). In singing, such correlations tend not to be automatic, as the very goal of many singing styles is to be different from speech (and as a contrastive feature in language, creak can be controlled separately from loudness).

[5]Singer 10 appears to sing with a singer's formant in every style. This was not achieved by training: this singer was a relative beginner, having had only 4 years of experience in professional folk singing. This singer is simply one of the rare individuals who have a spontaneous singer's formant.

[6]The presence of this original perceptual goal does not preclude the possibility that the "ring" or "brilliance" of the singer's formant has nowadays become a desirable perceptual goal of Western classical singing in itself, and is utilized as well outside the original conditions.

Bartholomew, W. T. (**1934**). "A physical definition of 'good voice-quality' in the male voice," J. Acoust. Soc. Am. **6**, 25–33.

Bergan, C. C., Titze, I. R., and Story, B. (**2004**). "The perception of two vocal qualities in a synthesized vocal utterance: ring and pressed voice," J. Voice **18**, 305–317.

Bezić, J. (**1968**). "Muzički folklor Sinjske krajine (The musical folklore of Sinjska Krajina)," Narodna Umjetnost—Croatian Journal of Ethnology and Folklore Research, Book **5–6**, 175–274.

Bezić, J. (**1979**). "Dalmatinske klapske pjesme kroz deset godina omiškog festivala (Dalmatian klapa songs throughout the ten years of the Omiš festival)," in *Zbornik dalmatinskih klapskih pjesama, izvedenih na festivalima u Omišu od 1967 do 1976 (Overview of Dalmatian klapa songs performed at the Omiš festival from 1967 to 1976)*, Omiš: Festival dalmatinskih klapa Omiš, pp. 16–23.

Bloothooft, G., and Plomp, R. (**1986**). "The sound level of the singer's formant in professional singing," J. Acoust. Soc. Am. **79**, 2028–2033.

Bloothooft, G., Bringmann, E., Van Cappellen, M., Van Luipen, J. B., and Thomassen, K. P. (**1992**). "Acoustics and perception of overtone singing," J. Acoust. Soc. Am. **92**, 1827–1836.

Boersma, P., and Weenink, D. (**2005**). "PRAAT: doing phonetics by computer (Version 4.3.31)" (Computer program), retrieved 27 November 2005, from http://www.praat.org.

Bonifačić, R. (**1996**). "*Tarankanje*: a disappearing music tradition," Narodna Umjetnost—Croatian Journal of Ethnology and Folklore Research **33**(1), 149–170.

Bonifačić, R. (**2001**). "O problematici takozvane 'istarske ljestvice' (On the topic of the so-called 'Istrian scale')," Narodna Umjetnost—Croatian Journal of Ethnology and Folklore Research **38**(2), 73–95.

Borch, D. Z., and Sundberg, J. (**2002**). "Spectral distribution of solo voice and accompaniment in pop music," Logoped. Phoniatr. Vocol. **27**, 37–41.

Burns, P. (**1986**). "Acoustical analysis of the underlying voice differences between two groups of professional singers: opera and country and western," Laryngoscope **96**, 549–554.

Ćaleta, J. (**1997**). "Klapa singing, a traditional folk phenomenon of Dalmatia," Narodna Umjetnost—Croatian Journal of Ethnology and Folklore Research **34**(1), 127–145.

Ćaleta, J. (**1999**). "The ethnomusicological approach to the concept of the Mediterranean in music of Croatia," Narodna Umjetnost—Croatian Journal of Ethnology and Folklore Research **36**(1), 183–195.

Cleveland, T. F., Sundberg, J., and Stone, R. E. (**2001**). "Long-term-average spectrum characteristics of country singers during speaking and singing," J. Voice **15**, 54–60.

Dobronić, A. (**1915**). "'Ojkanje,' prilog za proučavanje geneze naše pučke popijevke ('Ojkanje,' a contribution to the study of the origins of our folk songs)," in Zbornik za narodni život i običaje južnih Slavena (Proceedings on folk life and traditions of the Southern Slavs), Knjiga XX, svezak 1 (JAZU, Zagreb), pp. 1–25.

Doskov, D., Ivanov, T., and Boyanov, B. (**1995**). "Comparative analysis of singer's high formant in different type of singing voices," Folia Phoniatr Logop **47**, 291–295.

Estill, J., Fujimura, O., Erickson, D., Zhang, T., and Beechler, K. (**1994**). "Vocal tract contributions to voice qualities," in *Proceedings of the Stockholm Music Acoustics Conference (SMAC 93)*, edited by A. Friberg, J. Iwarsson, E. Jansson, and J. Sundberg (Royal Swedish Academy of Music, Stockholm), pp. 161–165.

Evans, L. B., Bass, H. E., and Sutherland, L. C. (**1971**). "Atmospheric absorption of sound: theoretical predictions," J. Acoust. Soc. Am. **51**, 1565–1575.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Fletcher, H., and Munson, W. A. (**1933**). "Loudness, its definition, measurement, and calculation," J. Acoust. Soc. Am. **5**, 82–108.

Frøkjær-Jensen, B., and Prytz, S. (**1976**). "Registration of voice quality," Brüel & Kjær Tech. Rev. **3**, 3–17.

Gauffin, J., and Sundberg, J. (**1989**). "Spectral correlates of glottal voice source waveform characteristics," J. Speech Hear. Res. **32**, 556–565.

Glave, R. D., and Rietveld, A. C. M. (**1975**). "Is the effort dependence of speech loudness explicable on the basis of acoustical cues," J. Acoust. Soc. Am. **58**, 875–879.

Hollien, H. (**1983**). "The puzzle of the singer's formant," in *Vocal Fold Physiology*, edited by D. M. Bless and J. H. Abbs (College-Hill, San Diego), pp. 368–378.

Johnson, A., Sundberg, J., and Wilbrand, H. (**1985**). "'Kölning.' Study of phonation and articulation in a type of Swedish herding song," in *Proceedings of the Stockholm Music Acoustics Conference (SMAC 83)*, edited by A. Askenfelt, S. Felicetti, E. Jansson, and J. Sundberg (Royal Swedish Academy of Music, Stockholm), pp. 187–202.

Karabaić, N. (**1956**). *Muzički folklor Hrvatskog primorja i Istre (Musical Folklore of the Croatian Littoral and Istria)* (Novi List, Rijeka).

Kiukaanniemi, H., Siponen, P., and Mattila, P. (**1982**). "Individual differences in the long-term speech spectrum," Folia Phoniatr. **34**, 21–28.

Klingholz, F. (**1993**). "Overtone singing: productive mechanisms and acoustic data," J. Voice **7**, 118–122.

Kuwabara, H., and Ohgushi, K. (**1984**). "Acoustic characteristics of professional male announcers' speech sounds," Acustica **55**, 233–240.

Leino, T. (**1994**). "Long-term average spectrum study on speaking voice quality in male actors," in *Proceedings of the Stockholm Music Acoustics Conference (SMAC 93)*, edited by A. Friberg, J. Iwarsson, E. Jansson, and

J. Sundberg (Royal Swedish Academy of Music, Stockholm), pp. 206–210.
Lindestad, P.-Å., Södersten, M., Merker, B., and Granqvist, S. (**2001**). "Voice source characteristics in Mongolian "throat singing" studied with high-speed imaging technique, acoustic spectra, and inverse filtering," J. Voice **15**, 78–85.
Löfqvist, A., and Mandersson, B. (**1987**). "Long-term average spectrum of speech and voice analysis," Folia Phoniatr. **39**, 221–229.
Marošević, G. (**1994**). "Ojkanje u izvandinarskim područjima Hrvatske (Ojkanje outside of the Dinaric region in Croatia)," Etnološka Tribina **17**, 91–102.
Marošević, G. (**2004**). "Traditional music," on *Croatian Folk Culture* (compact disc), edited by Z. Vitez (Institut za etnologiju i folkloristiku, Zagreb), pp. 408–419.
Nawka, T., Anders, L. C., Cebulla, M., and Zurakowski, D. (**1997**). "The speaker's formant in male voices," J. Voice **11**, 422–428.
Nolan, F. (**1983**). *The Phonetic Bases of Speaker Recognition* (Cambridge U.P., Cambridge, MA).
Nordenberg, M., and Sundberg, J. (**2003**). "Effect on LTAS of vocal loudness variation," in *Speech, Music and Hearing*, Q. Prog. Stat. Report 45 (Royal Institute of Technology, Stockholm), pp. 93–100.
Rapanić, Ž. (**1979**). "Deset godina Festivala dalmatinskih klapa (Ten years of the Festival of Dalmatian klapa)," in *Zbornik dalmatinskih klapskih pjesama, izvedenih na festivalima u Omišu od 1967 do 1976 (Overview of Dalmatian klapa songs performed at the Omiš festival from 1967 to 1976)*. Omiš: Festival dalmatinskih klapa Omiš, pp. 9–13.
Ross, J. (**1992**). "Formant frequencies in Estonian folk singing," J. Acoust. Soc. Am. **91**, 3532–3539.
Rossing, T. D., Sundberg, J., and Ternström, S. (**1986**). "Acoustic comparison of voice use in solo and choir singing," J. Acoust. Soc. Am. **79**, 1975–1981.
Schutte, H. K., and Miller, D. G. (**1993**). "Belting and pop, nonclassical approaches to the female middle voice: Some preliminary considerations," J. Voice **7**, 142–150.
Schutte, H. K., and Miller, R. (**1984**). "Resonance balance in register categories of the singing voice: a spectral analysis study," Folia Phoniatr. **36**, 289–295.
Sluijter, A. (**1995**). "Phonetic correlates of stress and accent," Ph.D. thesis, Leiden University (Holland Academic Graphics, The Hague).
Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT, Cambridge, MA).
Stone, R. E., Jr., Cleveland, T. F., and Sundberg, J. (**1999**). "Formant frequencies in country singers' speech and singing," J. Voice **13**, 161–167.
Stone, R. E., Jr., Cleveland, T. F., Sundberg, J., and Prokop, J. (**2003**). "Aerodynamic and acoustical measures of speech, operatic, and Broadway vocal styles in a professional female singer," J. Voice **17**, 283–297.
Sundberg, J. (**1972**). "A perceptual function of the 'singing formant'," in Speech Transmission Laboratory Quarterly Progress and Status Report **2–3**, 61–63.
Sundberg, J. (**1973**). "The source spectrum in professional singing," Folia Phoniatr. **25**, 71–90.
Sundberg, J. (**1974**). "Articulatory interpretation of the 'singing formant'," J. Acoust. Soc. Am. **55**, 838–844.
Sundberg, J. (**1995**). "Swedish voices in music," in Proceedings of the 13th International Congress of Phonetic Sciences, Stockholm, Vol. **1**, pp. 2–8.
Sundberg, J. (**2001**). "Level and center frequency of the singer's formant," J. Voice **15**, 176–186.
Sundberg, J., Andersson, M., and Hultqvist, C. (**1999a**). "Effects of subglottal pressure variation on professional baritone singers' voice sources," J. Acoust. Soc. Am. **105**, 1965–1971.
Sundberg, J., Fahlstedt, E., and Morell, A. (**2005**). "Effects on the glottal voice source of vocal loudness variation in untrained female and male voices," J. Acoust. Soc. Am. **117**, 879–885.
Sundberg, J., Cleveland, T. F., Stone, R. E., Jr., and Iwarsson, J. (**1999b**). "Voice source characteristics in six premier country singers," J. Voice **13**, 168–183.
Thalén, M., and Sundberg, J. (**2001**). "Describing different styles of singing: a comparison of a female singer's voice source in 'classical,' 'pop,' 'jazz' and 'blues,'," Logoped. Phoniatr. Vocol. **26**, 82–93.
Thunberg, G. C. (**2003**). "Spectral balance utilization in different speaking styles-a preliminary investigation," in Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, pp. 2149–2152.
Van Tongeren, M. C. (**2002**). *Overtone Singing—Physics and Metaphysics of Harmonics in East and West* (Fusica, Amsterdam).
Zwicker, E., and Feldtkeller, R. (**1967**). *Das Ohr als Nachrichtenempfänger* (Hirzel, Stuttgart).

Acoustic detection and classification of microchiroptera using machine learning: Lessons learned from automatic speech recognition

Mark D. Skowronski[a] and John G. Harris[b]
Computational Neuro-Engineering Lab, Electrical and Computer Engineering, University of Florida, Gainesville, Florida 32611

(Received 29 August 2005; revised 2 December 2005; accepted 26 December 2005)

Current automatic acoustic detection and classification of microchiroptera utilize global features of individual calls (i.e., duration, bandwidth, frequency extrema), an approach that stems from expert knowledge of call sonograms. This approach parallels the acoustic phonetic paradigm of human automatic speech recognition (ASR), which relied on expert knowledge to account for variations in canonical linguistic units. ASR research eventually shifted from acoustic phonetics to machine learning, primarily because of the superior ability of machine learning to account for signal variation. To compare machine learning with conventional methods of detection and classification, nearly 3000 search-phase calls were hand labeled from recordings of five species: *Pipistrellus bodenheimeri, Molossus molossus, Lasiurus borealis, L. cinereus semotus*, and *Tadarida brasiliensis*. The hand labels were used to train two machine learning models: a Gaussian mixture model (GMM) for detection and classification and a hidden Markov model (HMM) for classification. The GMM detector produced 4% error compared to 32% error for a baseline broadband energy detector, while the GMM and HMM classifiers produced errors of $0.6 \pm 0.2\%$ compared to $16.9 \pm 1.1\%$ error for a baseline discriminant function analysis classifier. The experiments showed that machine learning algorithms produced errors an order of magnitude smaller than those for conventional methods. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2166948]

PACS number(s): 43.80.Ev, 43.60.Uv [JAS] Pages: 1817–1833

I. INTRODUCTION

Since the early work of Griffin, who showed that bats use ultrasonic calls for hunting and navigation,[1] acoustic methods of studying bats have augmented traditional methods of observation and capture. Detection and identification of species are fundamental to bat research, and several studies have investigated the merits of acoustic methods.[2–4] The echolocation calls of bats are traditionally characterized by "global" features of the calls (e.g., duration and frequency extrema), measured either by hand or automatically. The call features are then compared with those from known species using a variety of techniques: discriminant function analysis,[5] decision trees,[6] neural networks,[7] synergetic algorithms,[8] or qualitative analysis.[9]

The conventional methods of automatic acoustic detection and classification of bat calls parallel the methods of automatic speech recognition (ASR) used in acoustic phonetics.[10,11] Both methods attempt to mimic expert readings of time-frequency plots. In acoustic phonetics, expert spectrogram readers assign phonetic labels to human speech by first segmenting the speech into broad categories (e.g., voiced, unvoiced, and silence) and then refining the segment endpoints using various acoustic measures. Acoustic phonetic ASR attempts to automate the techniques of acoustic phonetics used by expert spectrogram readers.[12]

However, acoustic phonetic ASR suffers from several limitations.[11] First, the method is based on subjective measures since phonetic endpoints are poorly defined and vary among observers. The inherent subjectivity of acoustic phonetics makes automated implementation difficult. Second, the method is sensitive to noise, and classification performance degrades rapidly compared to the performance of expert spectrogram readers as noise energy increases. The method relies on several acoustic measures (e.g., formant estimator, voice activity detector, and fricative measure) which vary in their noise robustness, and classification performance degrades according to the least robust measure. Third, acoustic phonetic ASR, as well as expert spectrogram readers, have difficulty accounting for the variability of speech. The rules used in acoustic phonetics are typically derived from read speech and generalize poorly to everyday conversational speech from a variety of speakers and speaking styles.

The limitations of acoustic phonetic ASR led researchers to explore other methods of ASR. Statistical models, such as the hidden Markov model, were first applied to ASR in the mid-1970s, and ASR research began to shift during the next decade from acoustic phonetics towards the machine learning paradigm.[13] The principal difference between the two paradigms of ASR is that acoustic phonetics is *expert driven* whereas machine learning is *data driven*. In machine learning ASR, features from short windows of time (20–40 ms) are extracted from speech, then the features are used to train

[a]Electronic mail: markskow@cnel.ufl.edu
[b]Electronic mail: harris@cnel.ufl.edu

TABLE I. List of species used in detection and classification experiments, along with the number of hand-labeled calls recorded for each species and location. See Table II for details of each recording system.

Species	No. of calls	System	Location
Pipistrellus bodenheimeri	601	Pettersson	Shore line, Dead Sea, Israel
Molossus molossus	291	Pettersson	St. Martin, West Indies
Lasiurus borealis	796	Avisoft	Pinery Provincial Park, Ontario, CA
Lasiurus cinereus semotus	601	Avisoft	Hawaii, USA
Tadarida brasiliensis	652	Matlab	Bat house, Gainesville, FL, USA

models of the speech signal. Acoustic phonetics relies on expert knowledge of the acoustic characteristics of speech to demarcate and classify the parts of the speech signal (phonemes or words), while machine learning relies on the validity of the models as well as the accuracy and robustness of the parameter estimation algorithms to do the same. The state-of-the-art in ASR uses machine learning methods because the features and classifiers are more robust to noise and are better at handling the variability found in human speech, compared to acoustic phonetic ASR.[11]

This paper demonstrates the superior abilities of machine learning methods of detection and classification of echolocation calls over conventional methods. Section II describes the data collection systems, details the method of determining hand-labeled calls from continuous recordings, and introduces the machine learning algorithms as well as the baseline algorithms used for detection and classification. The next section details the experiment results and shows examples depicting the performance differences of the various algorithms under study. Section IV discusses the experiment results and also the strengths and limitations of the machine learning paradigm. The final section summarizes the work.

II. METHOD

A. Data

Recordings from five species were made from separate locations using three recording setups. The first recording setup consisted of a Pettersson D980 bat detector (Pettersson Elektronik AB, Uppsala, Sweden) and a DAT tape recorder for storage and was used to record echolocation calls from the species *Pipistrellus bodenheimeri* and *Molossus molossus*. *P. bodenheimeri* was recorded along the southern shore of the Dead Sea in Israel while *M. molossus* was recorded in the West Indies. The output of the D980 was fed to a tape recorder for storage.

The second recording setup consisted of an Avisoft UltraSoundGate (USG) 116 recording system and laptop running Avisoft recorder software (Avisoft Bioacoustics, Berlin, Germany) and was used to record echolocation calls from the species *Lasiurus borealis* and *L. cinereus semotus*. *L. borealis* was recorded in Pinery Provincial Park near Grand Bend in southern Ontario, Canada, while *L. cinereus semotus* was recorded in the island chain of Hawaii, USA. The output of the USG was fed into the USB port of the laptop running the Avisoft recorder software for storage.

The third recording setup consisted of a custom system and was used to record echolocation calls from the species *Tadarida brasiliensis* outside a known roost located in north central Florida, USA. The custom system consisted of a Brüel & Kjær (B&K) 4939 1/4-in. microphone and B&K 2670 preamplifier, a custom-made portable amplifier and antialiasing filter (cutoff frequency: 95 kHz), a National Instruments 6036E DAQCard A/D PCMCIA card (200 kSamples/second, 16-bit resolution), and a laptop computer. The data were acquired from the A/D card and stored to the hard drive using a custom interface written for Matlab 6.5 (The Mathworks, Natick, MA, USA). The system allowed for continuous recordings, limited by battery life and hard drive space. Table I lists the species used in the following experiments as well as details of their recordings, and Table II lists a summary of the recording systems used in the following experiments.

B. Hand labels

To establish a ground truth for all recorded calls, individual calls were labeled by hand using a custom graphical user interface (GUI) in Matlab. A continuous recorded signal was divided into 0.55-s windows with 0.05 s of overlap to account for calls at the boundary of a window. After applying the noise reduction algorithm known as spectral mean subtraction (described below), a spectrogram of each window was displayed in the GUI along with an estimate for the frequency of maximum spectral magnitude for each analysis frame, which were 2.0 ms in duration with 80% overlap between frames. Therefore, endpoint resolution was 0.4 ms. Each frame was scaled by a Blackman window to reduce spectral leakage. The Blackman window, compared to the more-popular Hamming window, has a 50% wider main lobe, which increases spectral smearing, but a lower side lobe peak by 16 dB,[14] which reduces side lobe leakage and is desired in order to suppress echoes from higher frequencies which may overlap with the lower frequencies of a downward-sweeping frequency-modulated (FM) echolocation call. The GUI was capable of zooming in to individual calls to aid in visual endpoint determination. After labeling

TABLE II. Summary of recording systems used to collect data in detection and classification experiments. See Table I for a list of species and recording details.

System	Microphone	Amp	Resolution	fs, kHz	Storage
Matlab	B&K 4939 1/4″	Custom	16 bits	200	Harx drive
Pettersson	D980 detector	D980	16 bits	298	DAT tape recorder
Avisoft	USG 116 detector	USG 116	8 bits	250	Hard drive

all calls in a window, the calls were saved to disk, and the next consecutive window in the data was prepared for display. One minute of recorded audio required approximately 1 h to hand label all calls using the GUI, and endpoint hand labels for all species were determined over several sessions by one of the authors (MS). Approximately 35 min of recorded data were scanned for bat calls using the GUI during the hand-labeling process.

A key point about the process of hand-labeling calls employed for the current experiments was that no call, no matter how faint or fragmented, was arbitrarily excluded from use in the detection and classification experiments. Excluding calls that do not appear "good" would insert an upward bias into the experiment results, since those calls are expected to be the most difficult to detect or classify. However, we strove to include all calls that appeared to be search-phase calls. The very few terminal-phase calls, or feeding buzzes, encountered in the recordings were marked as calls for the purposes of detection but were excluded from the classification experiments.

C. Feature extraction

Features for all experiments were estimated over short, overlapping frames of the recorded signals. For the following experiments, frame duration was 1.6 ms, and the frame rate was 10 000 frames per second (0.1 ms between frames, or 94% overlap). The high frame rate allowed for reliable estimation of feature temporal derivatives, which were used by the machine learning algorithms.

For each frame, the peak energy E (also known as peak amplitude), in dB, and the frequency for the peak energy, F_0, in Hz, were estimated from a zero-padded fast Fourier transform (FFT) magnitude spectrum. For signals at all sampling rates used in the current experiments, the signal was zero-padded to a length of 2048 points, or roughly four times the original length, before calculating the FFT. The longer zero-padded FFT did not increase resolution of the original windowed signal, *per se*, but conveniently performed interpolation of the smeared power spectrum in order to produce finer frequency estimates. Log compression of the peak energy was used to make the distribution of peak energy more Gaussian.

The broadband energy (Parseval energy) of each frame was also calculated and used in the baseline detector, after normalizing out the noise floor. The noise floor for each recorded file was estimated by finding the mode of energy values over 2-ms non-overlapping rectangular windows over the entire file. Files were up to 62.5 s in length. Some recording sessions spanned several files, and noise floor estimates were determined for each file. The mode was estimated by finding the peak of the histogram of energy values, and the number of bins used to construct the histogram was 10% of the total number of available energy values for each file. The noise floor estimate was used to normalize the energy estimates used by the broadband energy-based detector in order to account for the differences in amplifier gain among the three recording systems used in the current experiments. Besides noise floor normalization, no other processing was performed on the signal prior to energy estimation. That is, the audible baseband was not removed through filtering, and no attempts were made to remove noise that may be present in the data. Filtering and other noise reduction techniques may improve the baseline performance, but such techniques are beyond the scope of the current work. Examples of the peak and broadband energy features are shown in Figs. 6(b) and 6(c), respectively.

1. Temporal derivatives

In addition to frequency and peak energy, the first and second temporal derivatives of both features were included as input for the machine learning algorithms in order to add slope and concavity information to the feature vector. Slope and concavity features, also known as Δ and $\Delta\Delta$ features, respectively, are common in ASR.[15] Furthermore, slope and concavity features, along with the frequency and peak energy features, provide the terms necessary to construct a local second-order polynomial curve fit (Taylor series) of a call's spectrogram. The curve fitting of simple functions to echolocation calls has been previously investigated in automated classification experiments[16,17] and in identifying individual bats,[18] and the features currently employed incorporate the advantages of curve fitting into the frame-based paradigm.

Temporal derivatives were calculated according to Eq. (1)

$$y'(p) = \sum_{m=-M}^{M} h(m)y(p-m), \tag{1}$$

where $y'(p)$ is the temporal derivative of the feature $y(p)$ (frequency or peak energy) for frame p and $h(m)$ is a non-causal differential filter of length $2M+1$. For differentiation, $h(m)=m$ for $m \in [-M, M]$.[15] The differential filter length M was chosen such that the duration of the filter was 0.5 ms, so for a frame rate of 10 000 frames per second, $M=2$.

2. Spectral mean subtraction

While normalization of the broadband energy estimate was achieved by subtracting off the noise floor, the other features used in the experiment were extracted after applying a noise reduction technique common in ASR. Spectral mean subtraction (SMS), similar to cepstral mean subtraction used in ASR,[19] was applied to the acoustic data in order to remove the recording environment transfer function of each data acquisition system. To implement SMS, the 1.6-ms frames of FFT power spectra were aligned in 0.55-s blocks to form a time-frequency spectrogram (approximately 5500 frames). The spectra were log compressed, then the temporal average of each FFT frequency bin was computed and subtracted for that bin across all frames. The log-compressed spectra were then transformed to magnitude spectra by exponential expansion. The technique of spectral mean subtraction is described by Eq. (2):

$$X(k,p) = \left| \sum_{n=0}^{N-1} x_p(n) e^{-j(2\pi/N)kn} \right|,$$

$$X_1(k,p) = \ln(X(k,p)),$$
$$X_2(k,p) = X_1(k,p) - \frac{1}{P}\sum_{t=1}^{P} X_1(k,t), \quad (2)$$
$$\hat{X}(k,p) = \exp(X_2(k,p)),$$

where $X(k,p)$ is the FFT magnitude spectrogram for FFT bin k and frame p, $x_p(n)$ is the time domain signal with discrete time n for the pth frame of length N samples, $X_1(k,p)$ is the log-compressed magnitude spectrogram, P is the number of frames used to calculate the temporal mean of $X_1(k,p)$, and $\hat{X}(k,p)$ is the magnitude spectrogram after applying spectral mean subtraction. SMS removes the recording environment transfer function from the recorded data, and any gain factor applied to the signal is normalized out as well.

Spectral mean subtraction stems from homomorphic filtering[14] and works as follows: a time domain signal is convolved with an unknown and undesired *linear* transfer function. For recordings of bat calls, the transfer function includes characteristics of the recording environment as well as the amplifier. The FFT in line 1 of Eq. (2) transforms convolution in the time domain to multiplication in the frequency domain, where the magnitude of the product of FFT spectra is equivalent to the product of the magnitude of FFT spectra. The log compression converts the multiplication of FFT magnitude spectra into addition of log magnitude spectra. The log magnitude spectrum of the transfer function is assumed to be constant over time, essentially contributing a constant offset to the log magnitude spectrum of the desired time domain signal. The constant offset for each FFT bin is removed by subtracting off the temporal mean of each bin of the log magnitude spectrum. The modified log magnitude spectrum is transformed back to the magnitude domain by exponential expansion. Figure 1 shows an example of spectral mean subtraction performed on a field recording of a faint yet distinct bat call.

D. Detection

For the detection experiment, time was discretized into 20-ms bins, and the detector determined whether or not at least one bat call occurred in some part of each bin. In this way, the events and outcomes were both discrete, which allowed for tabulation of the detection results in confusion matrices. The experiments were performed using three detector algorithms: a baseline algorithm based on broadband energy, a baseline algorithm based on power spectrum peak energy using SMS, and a machine learning algorithm known as a Gaussian mixture model (GMM) using SMS.

1. Baseline broadband energy detector

The baseline detector used short-term broadband energy, described above, and a threshold to detect calls from the background.[3,7] The frame energy was compared to a threshold to determine whether or not a 20-ms bin contained at least one call. If the maximum frame energy out of all frames in a bin was above the threshold, the bin was said to contain at least part of one bat call; otherwise, the bin was said to contain no calls. The threshold was swept to create a receiver operator characteristic (ROC) curve, the standard figure of merit for comparing various detectors.[20]

2. Baseline power spectrum peak energy detector

The broadband energy estimate of each frame used in the first baseline detector was replaced by the magnitude of the power spectrum peak calculated for each frame. SMS was applied to the 0.55-s block of frames prior to spectral peak estimation. As with the other baseline detector, a bin was said to contain at least part of one call if the maximum spectral peak was above a threshold. The threshold was swept to create a ROC curve for comparison with the other detectors.

3. Gaussian mixture model detector

The Gaussian mixture model is a statistical model for the probability density function (pdf) of a random variable and is expressed in Eq. (3):

$$p(\mathbf{x}) = \sum_{j=1}^{J} w(j) G(\mathbf{x}, \mu_j, \Sigma_j). \quad (3)$$

The term $p(\mathbf{x})$ is the pdf of the $N \times 1$ random variable $\mathbf{x}$, $G(\mathbf{x}, \mu_j, \Sigma_j)$ is the jth Gaussian kernel with an $N \times 1$ mean vector μ_j and $N \times N$ covariance matrix Σ_j, $w(j)$ is the Gaussian kernel weight for the jth kernel, and J is the number of Gaussian kernels (or mixtures) used in the model. For the automated detection and classification experiments, the random variable $\mathbf{x}$ was a vector of the $N=6$ features extracted from each analysis frame: peak energy E, frequency at peak energy F_0, and the temporal derivatives ΔE, ΔF_0, $\Delta\Delta E$, and $\Delta\Delta F_0$. The GMM has been shown to accurately estimate any continuous pdf, given enough kernels in the mixture.[21] The parameters for the GMM, μ_j and Σ_j and $w(j)$, were estimated in the standard way from hand-labeled training data using the expectation-maximization algorithm.[20]

Two Gaussian mixture models were trained as part of the GMM detector: one from the hand-labeled calls, and the other from the intercall acoustic background signal (including call echoes). The output of each GMM (the evaluation of the log likelihood $\log[p(\mathbf{x})]$ for a given feature vector $\mathbf{x}$) was computed, and the difference was compared to a threshold in a manner similar to that used with the baseline detectors. If any frame in a 20-ms bin produced a log likelihood from the call GMM larger than that from the background GMM by an amount above the threshold, the bin was labeled by the GMM detector as containing at least part of a bat call. Otherwise, the bin was said to contain no calls.

Temporal information was included in all detectors by adding an averaging window to the detector output. For the broadband energy-based detector and peak energy-based detector, the output for each analysis frame was energy in dB. For the GMM detector, the output was the log likelihood that a frame of features was produced by the pdf of features modeled by the mixture of Gaussian kernels. A Blackman win-

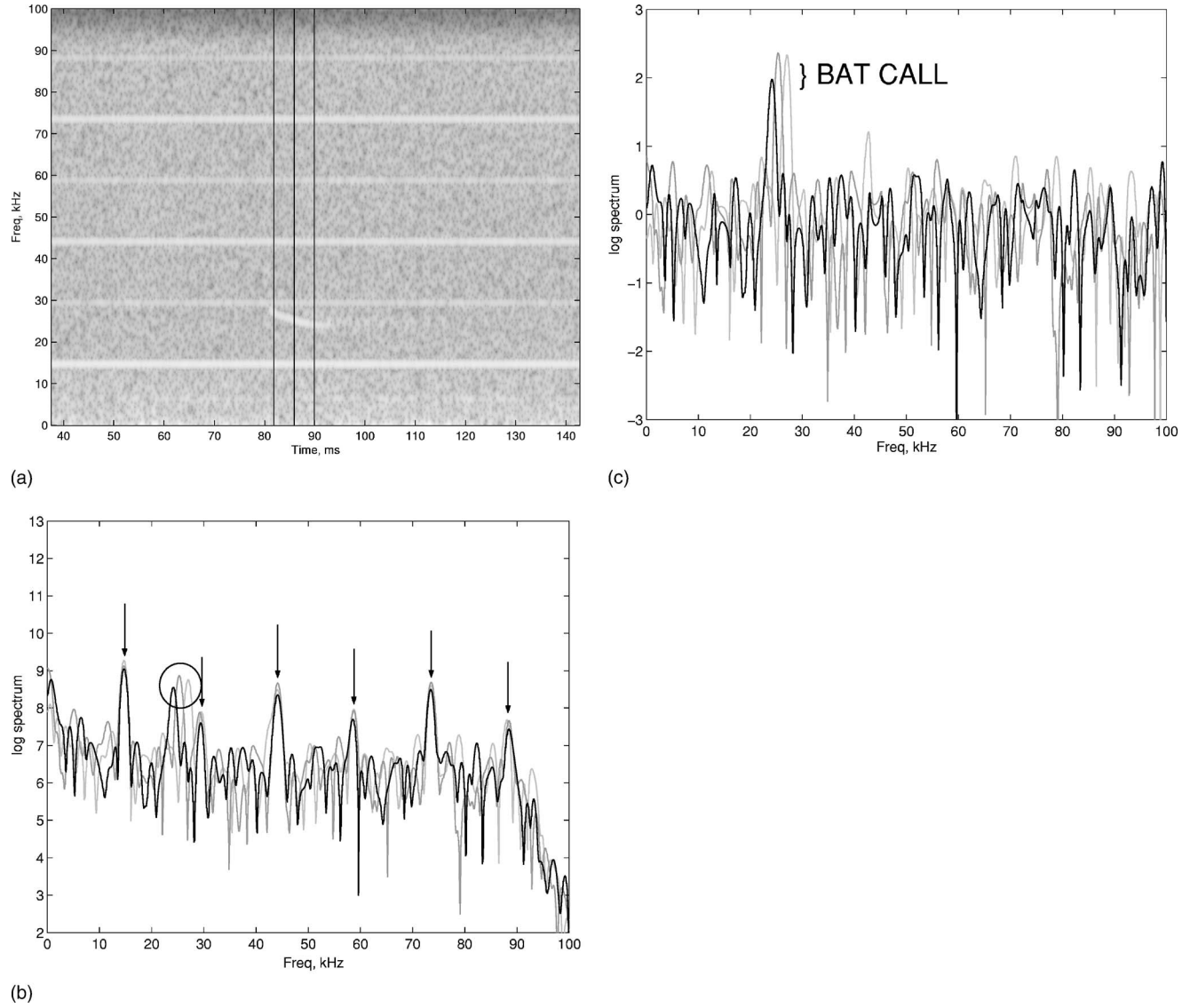


FIG. 1. Example of spectral mean subtraction performed on a faint bat call from *T. brasiliensis*. In (a), the spectrogram of the recorded call is shown before applying SMS, and darker shades indicate lower intensities. The three vertical black lines indicate the log spectra displayed before applying SMS, in (b), and after applying SMS, in (c). The darker lines in (b) and (c) represent windows later in time. In (b), the arrows point to harmonic peaks, generated by the switching power supply used to provide the 200-V polarization voltage to the microphone, and the circle highlights the bat call fundamental frequency peaks. In (c), the harmonics are suppressed by SMS, and the effects of low-pass filtering above 95 kHz as well as the spectral peak near 0 Hz are removed without degrading the echolocation call.

dow of length 5 ms was used to average the detector output and acted as a low-pass filter to temporally smooth the output.

Figure 2 shows the mixtures of Gaussian kernels for both the call and background detection GMMs, projected onto the features peak energy and frequency at peak energy, along with trajectories of the three hand-labeled calls shown in Fig. 3. Each Gaussian kernel in the mixture appears as a rotated and scaled ellipse in Fig. 2 according to that kernel's (full) covariance matrix Σ_j, and the kernels are located in feature space according to each kernel's mean vector μ_j. The center of each ellipse represents the maximum likelihood for that ellipse, and the edge of the ellipse represents the likelihood one standard deviation away from the peak.

The background GMM kernels in Fig. 2 are the light gray ellipses and are distributed across all frequencies at the low end of the peak amplitude range of values. The dark gray ellipses represent the call GMM kernels and are distributed across the relevant frequency and peak energy ranges for the species studied in the current experiments. The trajectories in Fig. 2 begin near the background level at frequencies near 50 kHz and sweep downward in frequency and upward in peak amplitude. The three trajectories plateau at different peak energy levels, due to the fact that the bat was approaching the recording microphone, as seen in Fig. 3(a), but all three trajectories terminate near 38 kHz and 5 dB. The three calls shown in Fig. 3 are part of the pass of 25 calls shown in Fig. 6. The trajectories overlap the kernels of the call GMM more so than the kernels of the background GMM, indicating that the trajectories are much more likely to be calls than part of the background [about 20 dB higher in log likelihood, according to Fig. 6(a)]. The frequency estimates shown in Fig. 3(b) appear as sonograms for the hand-labeled calls,

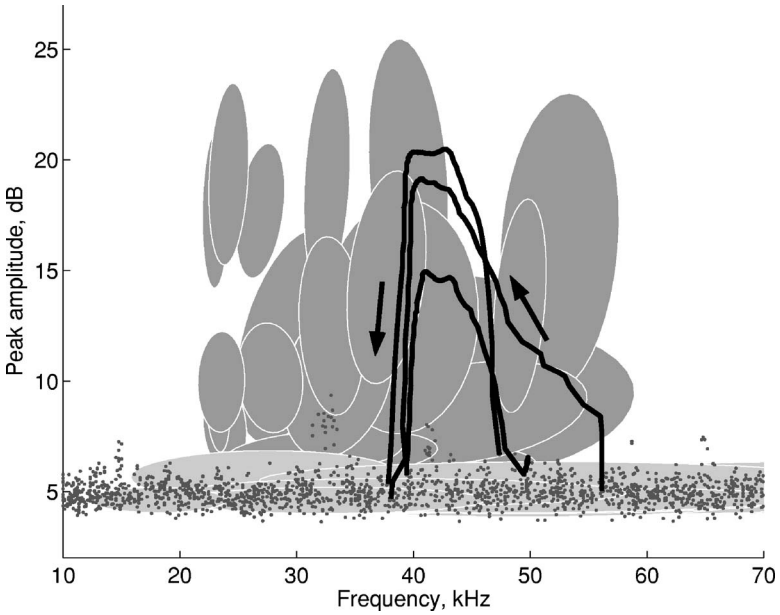


FIG. 2. Gaussian mixture model kernels for the call GMM (dark gray ellipses) and background GMM (light gray ellipses). Each ellipse is one Gaussian kernel in a GMM, projected onto the two features "peak amplitude" and "frequency at peak amplitude," and the axes for each ellipse represent one standard deviation for that particular kernel, according to its covariance matrix. The black lines are trajectories through feature space from the three hand-labeled calls in Fig. 3, and the gray dots are features from the intercall background. The arrows point in the direction along the trajectories that the features from the downward-sweeping calls traverse.

while the frequency estimates for the intercall background regions appear random and evenly distributed across the frequency range as expected.

E. Classification

To simplify the classification experiment, only hand-labeled calls were used for testing and training. Errors from the bat call detector would affect classifier performance, complicating the comparisons among the baseline and machine learning algorithms. The current results can be interpreted as representing classifier performance assuming ideal detection. Degradation in performance due to nonideal detection is beyond the scope of the current work.

1. Baseline classifier

The baseline classifier used discriminant function analysis (DFA) on the following features most commonly used in the literature: minimum frequency, maximum frequency, frequency of peak energy, and duration.[2,3,7–9,22] The frequency features were measured for each call using the frequency estimates used by the machine learning algorithms, which were made after applying SMS.

DFA is a statistical classifier used in many bat call classification experiments.[5,7,8,22,23] Furthermore, DFA is a standard linear classifier, included in many popular statistical software packages, and is readily available to many bat researchers. In the current experiments, DFA was implemented with the Statistics toolbox, version 4.0, in Matlab 6.5, using the settings "quadratic" and "empirical" for the parameters "type" and "prior," respectively. The quadratic type of DFA means that each species was modeled with a separate (i.e., stratified) covariance matrix, and the empirical prior setting instructed the DFA algorithm to estimate the *a priori* probability for each species using the training data.

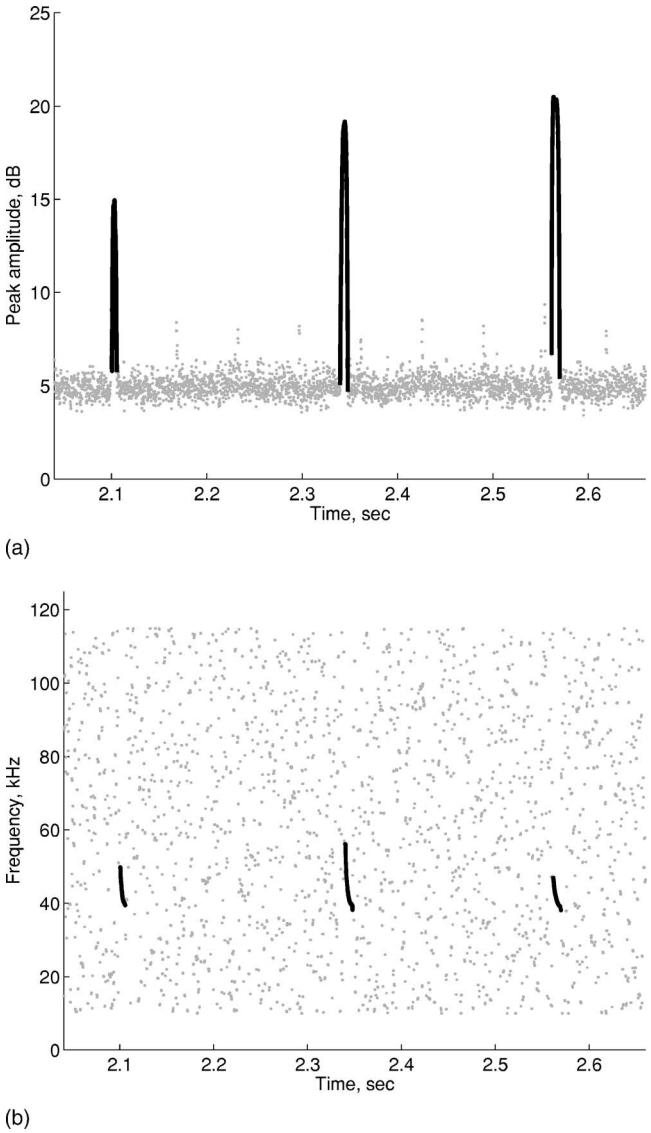


FIG. 3. Features from three hand-labeled calls of *Lasiurus borealis*: peak amplitude in (a) and frequency at peak amplitude in (b). The black curves are for hand-labeled calls, while the gray dots are features from the background. The features from the hand-labeled calls form the three trajectories through feature space shown in Fig. 2.

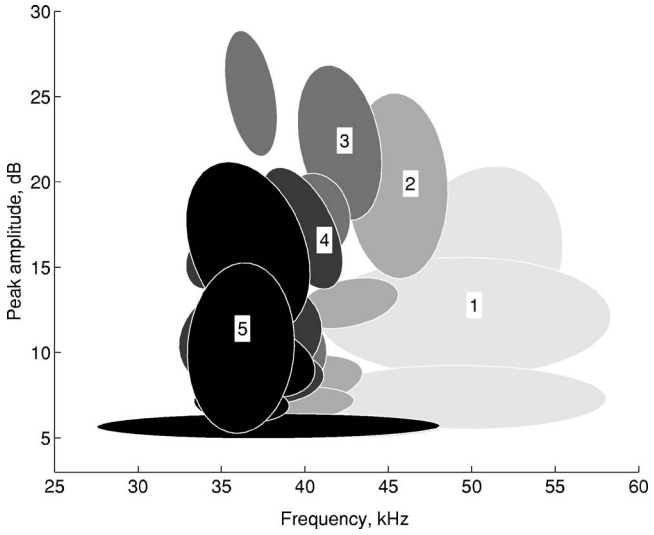


FIG. 4. Gaussian kernels for the HMM classifier for *Lasiurus borealis*, similar to the GMM detector kernels shown in Fig. 2. The states are numbered from 1 (first) to 5 (last), and each shade of gray corresponds to a state in the model. Each state is represented by five Gaussian kernels (ellipses). Calls typically begin in state 1, although some may begin in state 2, and transition through the states sequentially until eventually terminating in state 5.

2. Gaussian mixture model classifier

The input to the GMM classifier was the same as the input to the GMM detector. Features, including derivatives, were extracted *before* the calls were excised from the background using the hand labels, thus preserving the boundaries of the feature vector sequences. For the GMM detector, one model was trained using bat calls from all species. For classification, however, one model was trained for each species of bat. Since only hand-labeled calls were used to train the GMMs, no model for the background was necessary. The output of each model was the averaged log likelihood that an input sequence of feature vectors was generated by that model. An unknown test sequence was classified as the species whose model produced the highest log likelihood. All species were assumed to have equal *a priori* probability of occurrence, although the number of calls labeled for each species, listed in Table I, varied from 300 to 800 calls.

3. Hidden Markov model classifier

The second machine learning algorithm used for classification was the hidden Markov model (HMM).[13] A HMM is similar to a GMM in that both methods model the probability density functions of features extracted from a time series. However, a HMM includes a temporal constraint that considers the production of a nonstationary signal as a piecewise-stationary process. The pdf of each stationary region of a signal is typically modeled in a HMM by a mixture of Gaussian kernels, where each stationary region is referred to as a *state*. A GMM is, in effect, a one-state HMM. The temporal constraint is achieved in a HMM by weighting the transitions between states through a *state transition matrix*.

Figure 4 shows the Gaussian kernels for each state in the HMM for the species *Lasiurus borealis*, similar to the GMM kernels (for all species) shown in Fig. 2. Each of the five states is represented by a shade of gray—lightest for state 1 and darkest for state 5—and each state is represented by five Gaussian kernels (ellipses). Some calls from *Lasiurus borealis* start in state 1 between 45 and 55 kHz and from 5 to 20 dB in peak amplitude; however, state 2 in Fig. 4 extends from near the noise floor at 5-dB peak amplitude to 25 dB and from 40 to 45 kHz, and some calls skip state 1 and start in state 2 near the noise floor before traversing to state 3 where the peak energy of each call is modeled. Figure 9 shows that some calls of *Lasiurus borealis* have a maximum frequency near 50 kHz while another group of calls start at a lower maximum frequency near 40 kHz. The multimodal distribution of maximum frequency, seen as distinct Gaussian kernels near the noise floor for states 1 and 2 in Fig. 4, is easily handled by the HMM but poorly modeled by a DFA classifier which assumes a unimodal distribution of features.

For the following experiments, the parameters of the state pdf models and the transition probabilities in the state transition matrix were learned from hand-labeled training data using the Baum-Welch algorithm. The output of a HMM was the log likelihood that a sequence of feature vectors was generated by a HMM along the single most likely sequence of states and was determined using the Viterbi algorithm. Both the Baum-Welch and Viterbi algorithms are standard for training and testing a HMM, respectively.[13]

A left-to-right HMM architecture of states was used to model the temporal progression of bat calls. A call typically begins in the first, or leftmost, state of an HMM, progressing through the sequence of states towards the final, or rightmost, state at the end of a call. An important feature of the HMM for modeling bat calls is that, for a call fragment, the call may begin after the first state and end before the final state. Call fragments are pervasive in echolocation recordings, due to the natural variability present in the production of echolocation calls and also due to the limited range of detection of calls because of the noise floor of the recording system. The ability of the HMM to account for fragmented calls is an advantage of the model over the baseline DFA classifier.

The input to the HMM classifier was the same as that for the GMM detector and classifier. As with the GMM classifier, one HMM was trained for each species of bat using hand-labeled calls. The output of each model was the log likelihood that the input sequence was generated by each HMM along the Viterbi sequence, the single most likely sequence of states through the model. An unknown test sequence was labeled as the species whose HMM produced the highest log likelihood, assuming equal *a priori* probability for each species.

F. Model parameters

For both the GMM and HMM, the models contain parameters which must be either learned from the data or set by the experimenter. For the current experiments, both models contained parameters for the probability density functions of extracted features, and each kernel in the pdf model was represented by a mean vector and a full covariance matrix.

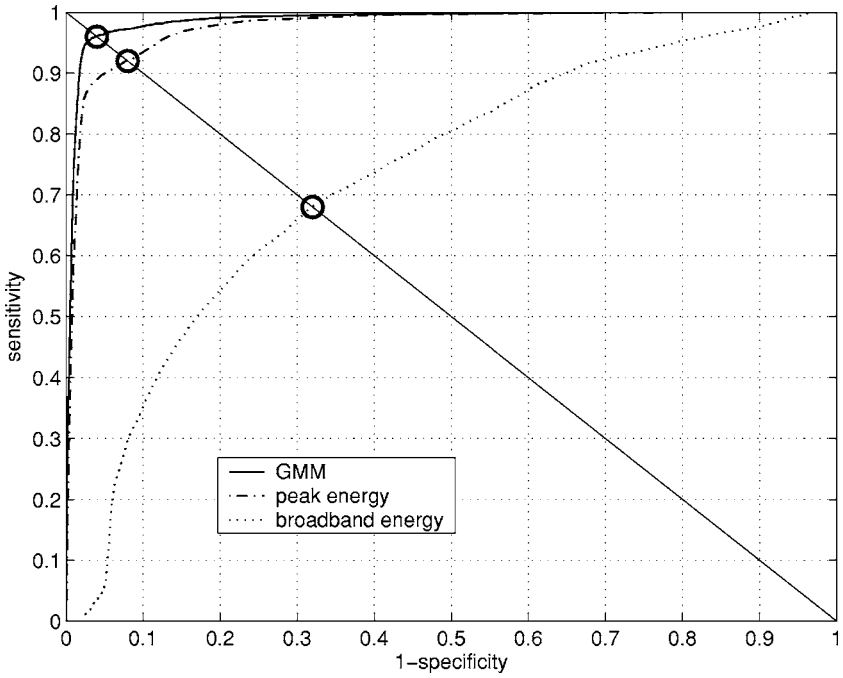


FIG. 5. Receiver operator characteristic curves for the resubstitution detection experiment. The abscissa is a fraction of false positives (i.e., 1—specificity) and the ordinate is a fraction of true positives (i.e., sensitivity). An ideal detector would have an operating point in the upper left corner of the plot (sensitivity and specificity of 1). The circles on the ROC curves indicate the operating points with sensitivity=specificity, and the confusion matrices for the circled points are shown in Table III. The area under the curves (AUC) are 0.988, 0.977, and 0.705 for the GMM, peak energy, and broadband energy detectors, respectively.

The pdf model also included a kernel weighting vector, and all model parameters were estimated from the hand-labeled data using expectation-maximization algorithms. All model training and testing were performed using the Full Bayes Network Toolbox[24] and Matlab 6.5.

Besides pdf mean and covariance parameters, the number of kernels used in each pdf model must be set by the experimenter. Any continuous pdf can be represented to a given accuracy tolerance using a GMM with a sufficient number of kernels, although the number of parameters to estimate for a GMM is proportional to the number of kernels. For a finite amount of data, the variance of each estimated parameter increases as the number of kernels increases. Thus, the experimenter must choose the number of kernels based on the tradeoff between pdf accuracy and parameter variance. For the detection and classification experiments, a total of 25 kernels were used for the pdf models in both the GMM and HMM. For the HMM, five states were used in the model, and each state used five kernels. As the experiment results below show, the number of kernels chosen, although not optimized, produced models of high accuracy relative to the baseline. For classification, the nearly identical results for a resubstitution experiment and cross-validation experiment show that the pdf models generalized well to novel data, validating the choice of the number of kernels.

III. RESULTS

A. Detection

To train the call GMM and background GMM for the GMM detector, all hand-labeled calls and 2% of all intercall background segments, selected uniformly randomly from all available data, were used. The percentage of background segments, used was smaller than that for calls because the background segments were typically much longer in duration. Approximately 250 k and 500 k frames were used to train the call and background models, respectively.

To create the ROC curves for each detector, minimum and maximum thresholds appropriate for each detector were determined (producing zero false negatives and zero false positives, respectively), then 2000 linearly spaced thresholds between the minimum and maximum thresholds were selected. The detectors were evaluated by comparing hand labels for all 0.55-s windows with detector output in order to determine which 20-ms bins in each window contained part of at least one call. Because both testing and training used all of the available calls, the resubstitution experiment results represent a "best case" estimate of generalized detection experiment results using cross-validation with novel test data. However, the classification results below show that the resubstitution classification results nearly match the cross-validation results, indicating that the machine learning algorithms generalize well using the current training methods and data set.

Figure 5 shows the ROC curves for each of the three detectors used in the resubstitution detection experiment. The circles on the ROC curves indicate the operating points with equal sensitivity and specificity of the detectors, and the confusion matrices for those operating points are shown in Table III. Table III lists several performance measures for the three detectors, including positive predictive value (PPV, number of true positives divided by the number of true and false positives) and negative predictive value (NPV, number of true negatives divided by the number of true and false negatives).

The ROC curves in Fig. 5 show significant differences in performance between the broadband-energy-based baseline and the other detectors. Without accounting for baseband noise or the recording environment, the broadband-energy-based detector produces detection errors an order of magnitude greater than the GMM detector. The peak-energy-based detector accounts for the recording environment by using SMS and employs a noise reduction algorithm by es-

TABLE III. Confusion matrices for the operating points of each ROC curve circled in Fig. 5. The rows are labeled as the ground truth while the columns are labeled as the detector output. CM=confusion matrix, T=true, F=false, Sens.=sensitivity, Spec.=specificity, PPV=positive predictive value, and NPV=negative predictive value.

		CM						
Detector		T	F	Sens.	Spec.	PPV	NPV	Error
GMM	T	4314	176	0.96	0.96	0.52	0.998	0.040
	F	3985	96 615					
Peak	T	4132	358	0.92	0.92	0.35	0.996	0.078
energy	F	7819	92781					
Broadband	T	3047	1443	0.68	0.68	0.087	0.98	0.32
energy	F	31 891	68 709					

sentially filtering out all noise not in the spectral vicinity of the power spectrum peak. Using both techniques, the peak-energy-based detector error was significantly lower than the broadband-energy-based detector error, although the error was still larger by about a factor of 2 compared to the GMM detector. The performance gains of the GMM detector over the peak-energy-based detector are due to the fact that the GMM detector used more information from the calls *and* the background. An example of detector performance for a single pass of calls is shown in Fig. 6. The pass contains 25 hand-labeled calls, denoted by gray bars in the figure. Figure 6 demonstrates how the broadband energy detector distinguishes only the most prominent calls, near the middle of the pass, from the background noise, while the GMM detector output peaks significantly above the background output level at the locations of all of the hand-labeled calls.

B. Classification

Two classification experiments were performed: (1) a resubstitution experiment to generate a classification confusion matrix for the five species used in the experiment to see where the classification errors were made and (2) a cross-validation experiment to quantify the generalization performance of each classifier to novel test data and also to compare with the optimistic resubstitution experiment results. For the resubstitution experiment, all of the available hand-labeled bat calls were used to train the models in the machine learning and DFA algorithms, one model for each species. For each trial of the cross-validation experiment, 50% of all calls were uniformly randomly selected for training the models for each of the three classifiers while the remaining 50% of the calls were used to test the classifiers. That is, all three classifiers were trained on the same calls for each trial, and no call used in training was also used in testing.

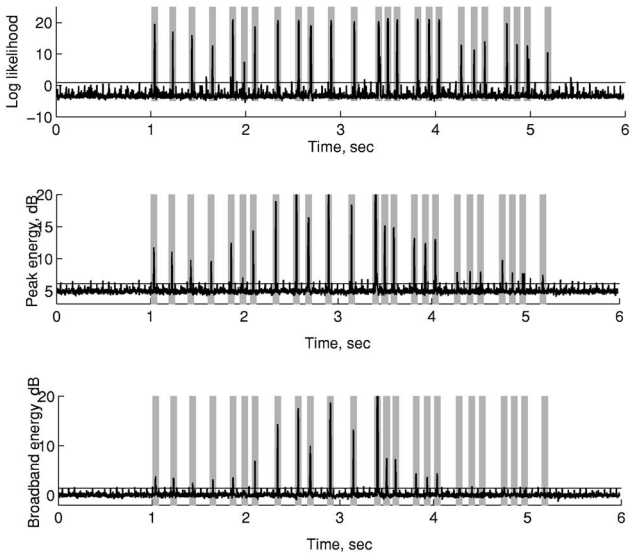


FIG. 6. Detector outputs for a single pass of 25 hand-labeled calls from *Lasiurus borealis* using (a) the GMM detector, (b) the peak-energy-based detector, and (c) the broadband-energy-based detector. The gray bars denote the locations of hand-labeled calls, and the horizontal black lines denote the thresholds for equal sensitivity and specificity, denoted by the circles in Fig. 5.

TABLE IV. Confusion matrix for the HMM classifier from the resubstitution experiment. Values in each cell are the number of calls from the hand-labeled species for each row that were classified as the species for each column. The species are *Pb: Pipistrellus bodenheimeri, Mm: Molossus molossus, Lb: Lasiurus borealis, Lc: Lasiurus cinereus semotus*, and *Tb: Tadarida brasiliensis*.

Species	*Pb*	*Mm*	*Lb*	*Lc*	*Tb*	% Correct
Pb	599	0	0	0	2	99.7
Mm	0	283	0	0	8	97.3
Lb	0	0	796	0	0	100
Lc	0	0	0	601	0	100
Tb	0	0	0	0	652	100
						99.7

1. Resubstitution experiment

The resubstitution classification results are reported as the confusion matrices in Tables IV–VI for the HMM, GMM, and DFA classifiers, respectively. The rows of each confusion matrix are labeled with the ground truth, while the columns are labeled with the classifier output. Each table entry a_{ij} is the number of calls a from the hand-labeled species in row i that were classified as the species in column j. The last column in each table is the percent correct classification for each species, while the value in the lower right-hand corner of each table is the overall percent correct classification for the given classifier.

TABLE V. Confusion matrix for the GMM classifier from the resubstitution experiment. Values in each cell are the number of calls from the hand-labeled species for each row that were classified as the species for each column. The species are *Pb: Pipistrellus bodenheimeri, Mm: Molossus molossus, Lb: Lasiurus borealis, Lc: Lasiurus cinereus semotus*, and *Tb: Tadarida brasiliensis*.

Species	*Pb*	*Mm*	*Lb*	*Lc*	*Tb*	% Correct
Pb	601	0	0	0	0	100
Mm	0	282	0	0	9	96.9
Lb	0	0	795	1	0	99.9
Lc	0	0	0	601	0	100
Tb	0	1	0	0	651	99.8
						99.6

TABLE VI. Confusion matrix for the DFA classifier from the resubstitution experiment. Values in each cell are the number of calls from the hand-labeled species for each row that were classified as the species for each column. The species are *Pb: Pipistrellus bodenheimeri, Mm: Molossus molossus, Lb: Lasiurus borealis, Lc: Lasiurus cinereus semotus,* and *Tb: Tadarida brasiliensis*.

Species	*Pb*	*Mm*	*Lb*	*Lc*	*Tb*	% Correct
Pb	581	1	19	0	0	96.7
Mm	2	226	8	51	4	77.7
Lb	7	138	631	2	18	79.3
Lc	0	3	3	552	43	91.8
Tb	0	39	39	117	457	70.1
						83.2

TABLE VIII. Confusion matrix for the GMM classifier from the cross-validation experiment. Values in each cell are the average percentage (±1 standard deviation) of calls from the hand-labeled species for each row that were classified as the species for each column over 20 trials. The species are *Pb: Pipistrellus bodenheimeri, Mm: Molossus molossus, Lb: Lasiurus borealis, Lc: Lasiurus cinereus semotus*, and *Tb: Tadarida brasiliensis*.

Species	*Pb*	*Mm*	*Lb*	*Lc*	*Tb*	% Correct
Pb	99.6±0.4	0±0	0±0	0±0	0.4±0.4	99.6±0.4
Mm	0.03±0.1	96.2±1.4	0±0	0±0	3.7±1.4	96.2±1.4
Lb	0±0	0±0	99.8±0.2	0.2±0.2	0±0	99.8±0.2
Lc	0±0	0±0	0.2±0.2	99.8±0.2	0±0	99.8±0.2
Tb	0±0	0.2±0.3	0±0	0±0	99.8±0.3	99.8±0.3
						99.4±0.2

2. Cross-validation experiment

The second classification experiment examined the generalization of the classification models to novel test calls. For each of 20 trials, 50% of all calls were uniformly randomly selected to train the three sets of models (HMM, GMM, and DFA) while the remaining 50% of the calls were used to test the classifiers.

The cross-validation classification results are reported as the confusion matrices in Tables VII–IX for the HMM, GMM, and DFA classifiers, respectively. As with Tables IV–VI, the rows of each confusion matrix are labeled with the ground truth, while the columns are labeled with the classifier output. Each table entry a_{ij} is the average percentage of calls a from the hand-labeled species in row i that were classified as the species in column j, ±1 standard deviation. The average and standard deviation were calculated over all trials. The last column in each table is the average percent correct classification for each species, while the value in the lower right-hand corner of each table is the overall average percent correct classification for the given classifier. For both the HMM and GMM classifiers, classification accuracy was 99.4±0.2% correct; for the baseline DFA classifier, classification was 83.1±1.1% correct. A t test between the GMM and HMM overall average scores across all trials showed that the scores were not significantly different ($p > 0.9$).

Figures 7–11 show example sonograms for each of the five species represented in the classification experiment. In the top panel of each figure, frequency estimates (i.e., sonograms) for 15 randomly selected hand-labeled calls from one of the species are shown. A large "X" appears behind a set of probabilities that produced an incorrect classification, and the calls are artificially spaced by 4-ms gaps for display purposes. The other three panels in each figure show the *a posteriori* probabilities for each detector using the accompanying call from the first panel as input. The *a posteriori* probability is determined from the *a priori* probability and pdf using the Bayes rule:[20]

$$p(\lambda_i|\mathbf{x}) = \frac{p(\mathbf{x}|\lambda_i)P(\lambda_i)}{\sum_{\forall i} p(\mathbf{x}|\lambda_i)P(\lambda_i)}, \tag{4}$$

where $p(\lambda_i|\mathbf{x})$ is the *a posteriori* probability that feature vector $\mathbf{x}$ came from the species λ_i, $p(\mathbf{x}|\lambda_i)$ is the (conditional) pdf from Eq. (3) for λ_i, and $P(\lambda_i)$ is the *a priori* probability for λ_i. The denominator in Eq. (4) is a scale factor that ensures that $p(\lambda_i|\mathbf{x})$ is a proper probability such that $\Sigma_{\forall i} p(\lambda_i|\mathbf{x}) = 1$.

IV. DISCUSSION

The results from the machine learning algorithms compared to the baseline detector and classifier show significant improvements in performance over the existing technology. The reasons for improved performance include the noise-reduction algorithm SMS, the use of more information from the audio signal, and more accurate statistical models.

TABLE VII. Confusion matrix for the HMM classifier from the cross-validation experiment. Values in each cell are the average percentage (±1 standard deviation) of calls from the hand-labeled species for each row that were classified as the species for each column over 20 trials. The species are *Pb: Pipistrellus bodenheimeri, Mm: Molossus molossus, Lb: Lasiurus borealis, Lc: Lasiurus cinereus semotus*, and *Tb: Tadarida brasiliensis*.

Species	*Pb*	*Mm*	*Lb*	*Lc*	*Tb*	% Correct
Pb	99.8±0.3	0±0	0±0	0±0	0.2±0.3	99.8±0.3
Mm	0.03±0.2	95.6±1.6	0±0	0±0	4.3±1.6	95.6±1.6
Lb	0±0	0±0	99.8±0.1	0.2±0.1	0±0	99.8±0.1
Lc	0±0	0±0	0.2±0.2	99.8±0.2	0±0	99.8±0.2
Tb	0±0	0.2±0.2	0±0	0±0	99.8±0.2	99.8±0.2
						99.4±0.2

TABLE IX. Confusion matrix for the DFA classifier from the cross-validation experiment. Values in each cell are the average percentage (±1 standard deviation) of calls from the hand-labeled species for each row that were classified as the species for each column over 20 trials. The species are *Pb: Pipistrellus bodenheimeri, Mm: Molossus molossus, Lb: Lasiurus borealis, Lc: Lasiurus cinereus semotus*, and *Tb: Tadarida brasiliensis*.

Species	*Pb*	*Mm*	*Lb*	*Lc*	*Tb*	% Correct
Pb	97.1±0.8	0.2±0.2	2.7±0.8	0±0	0±0	97.1±0.8
Mm	0.6±0.5	76.7±3.2	4.1±2.3	17.3±2.7	1.3±0.6	76.7±3.2
Lb	1.2±0.4	16.9±1.5	79.6±1.3	0.3±0.3	2.1±0.5	79.6±1.3
Lc	0±0	1.1±0.9	0.3±0.5	89.7±1.4	8.8±0.9	89.7±1.4
Tb	0±0	6.6±1.5	5.4±1.4	16.5±2.6	71.4±3.3	71.4±3.3
						83.1±1.1

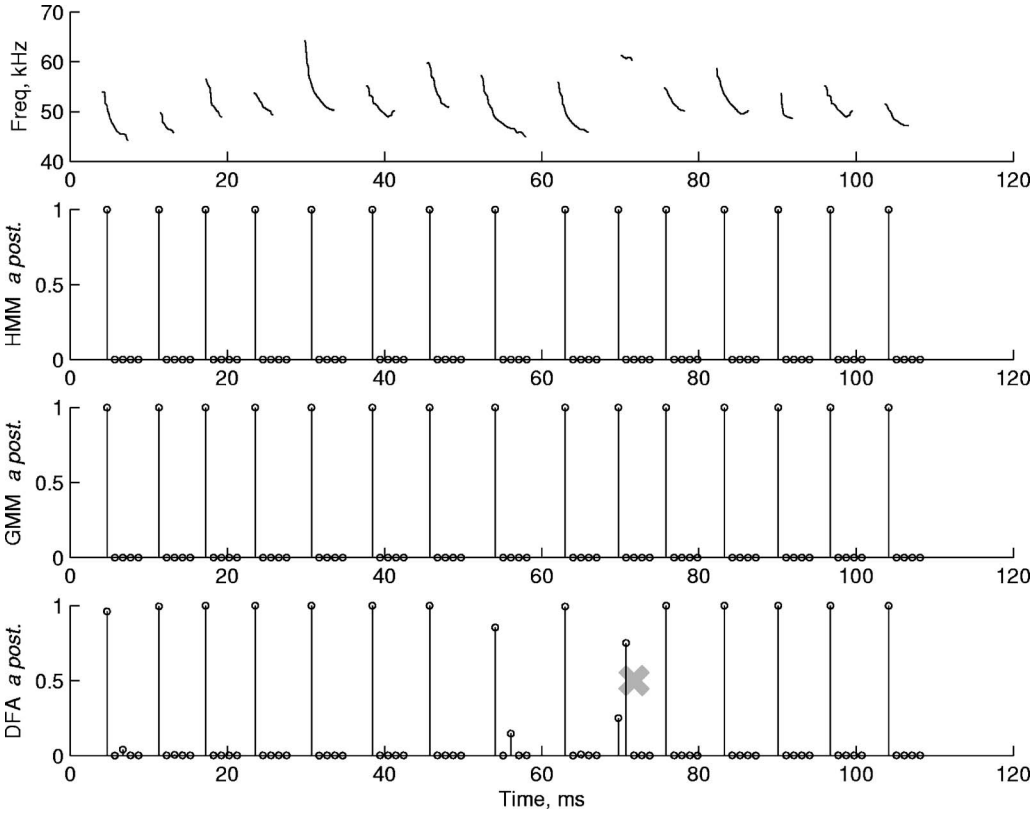


FIG. 7. Example sonograms and detector *a posteriori* probabilities for 15 randomly selected calls of *P. bodenheimeri.* Each call and corresponding groups of output probabilities from each detector are aligned vertically. The order of output probabilities for each call, left to right, is *P. bodenheimeri, M. molossus, L. borealis, L. cinereus semotus*, and *T. brasiliensis.* An "X" behind a group of probabilities indicates incorrect classification.

A. Feature extraction

All audio signals were preprocessed with SMS to remove the variations among recording environments and also to reduce harmonic amplifier noise present in several recordings. SMS is a simple algorithm, implemented in Eq. (2), which requires no parameters to set and makes no assumptions about the data other than the stationarity of the noise and recording environment transfer function. The related al-

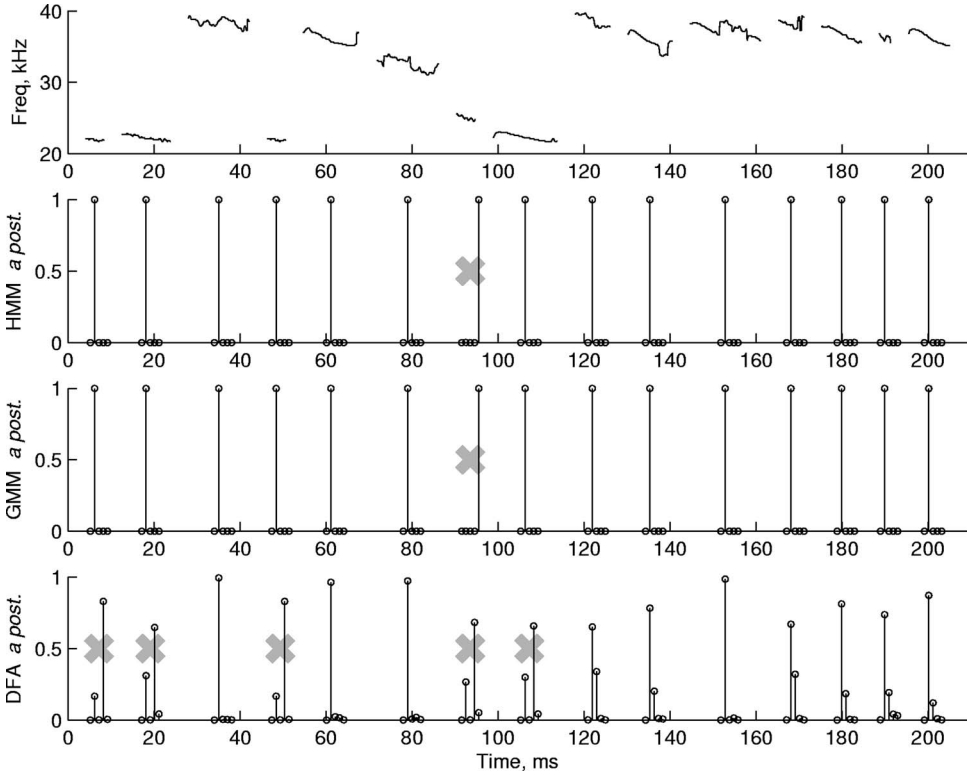


FIG. 8. Example sonograms and detector *a posteriori* probabilities for 15 randomly selected calls of *M. molossus*. Each call and corresponding groups of output probabilities from each detector are aligned vertically. The order of output probabilities for each call, left to right, is *P. bodenheimeri, M. molossus, L. borealis, L. cinereus semotus*, and *T. brasiliensis*. An "X" behind a group of probabilities indicates incorrect classification.

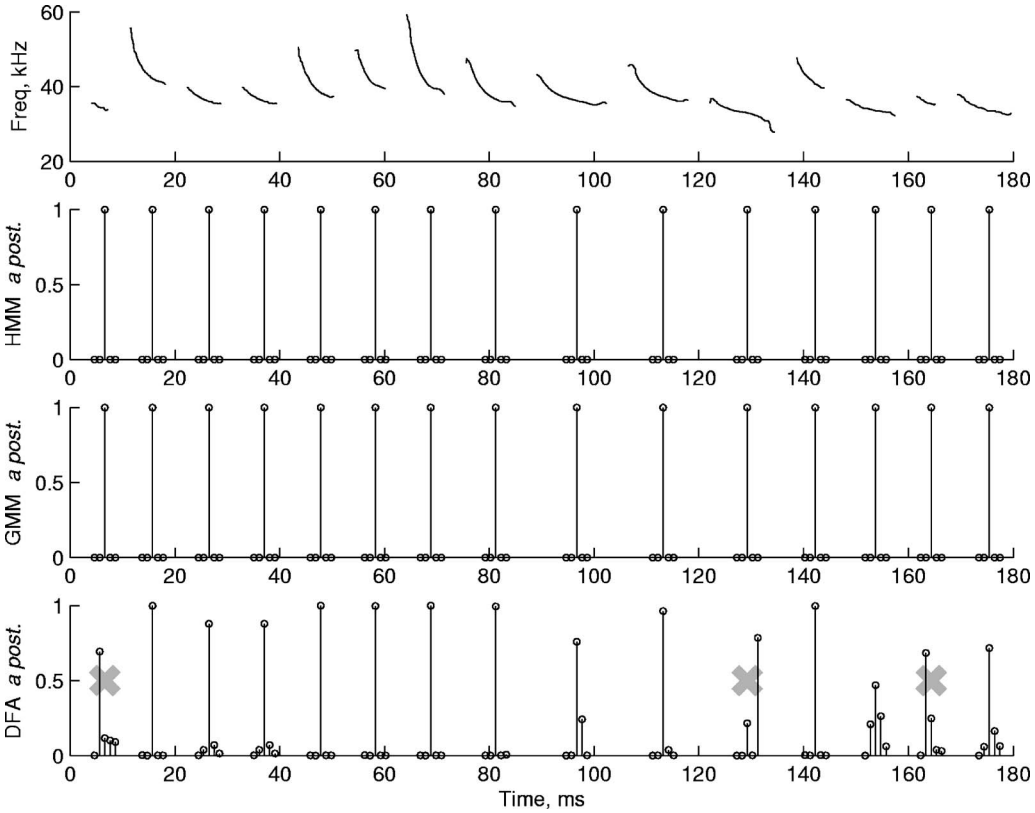


FIG. 9. Example sonograms and detector *a posteriori* probabilities for 15 randomly selected calls of *L. borealis*. Each call and corresponding groups of output probabilities from each detector are aligned vertically. The order of output probabilities for each call, left to right, is *P. bodenheimeri, M. molossus, L. borealis, L. cinereus semotus*, and *T. brasiliensis*. An "X" behind a group of probabilities indicates incorrect classification.

gorithm called cepstral mean subtraction, used extensively in ASR experiments since its inception, has maintained its popularity because of its simplicity, low computational cost, and effectiveness.

After applying SMS, frame-based features were extracted with the assumption that the relevant spectral information present in echolocation calls is contained in a single harmonic peak. By contrast, the relevant spectral feature of human speech containing phonetic information is the vocal tract response, which is superimposed with glottal harmonics for voiced speech. For bats, however, too few harmonics are typically recorded to reliably estimate the vocal tract. Thus, pitch estimation is the primary spectral characteristic from which to detect and classify bats. Several high-resolution methods exist for estimating the harmonic content of a signal corrupted by noise, such as MUSIC or ESPRIT.[25] However, the high-resolution methods are iterative and slow, so for simplicity and efficiency we chose to estimate fundamental frequency using a zero-padded FFT. For frames within a call, the peak energy of the spectrum corresponded to the fundamental frequency, or rarely the first harmonic, of the call, while for frames containing only background signal, the peak energy was nearly uniformly randomly distributed across all frequencies in the spectrum. SMS was effective at removing harmonic amplifier noise which would otherwise interfere in the peak energy estimation, as seen in Fig. 1.

The energy and peak harmonic features were augmented using first- and second-order temporal derivatives, a practice common in ASR. The derivatives provide an embedding of temporal information into each feature vector. For signals, such as bat calls, that have adjacent frames that are highly correlated, embedding temporal information increases the distinction between the frames of bat calls with frames of background noise, which are typically uncorrelated. Furthermore, temporal derivatives aid the machine learning algorithms in distinguishing up-sweeping from down-sweeping FM bat calls.

The current frequency estimation method has difficulty distinguishing calls that overlap in time since only a single peak frequency is estimated for each analysis frame. However, the method could be extended to identify multiple spectral peaks, either from temporally overlapping calls or harmonics of the fundamental frequency of a single call. One difficulty in extending the method is that the number of local peaks to identify is generally unknown and would have to be estimated.

B. Detection

The GMM detector reduced detection errors by an order of magnitude compared to the baseline broadband energy detector, according to the results in Table III. The resubstitution experiment involved about 105 000 20-ms bins, with about 4% of the bins containing part of at least one bat call according to the hand labels. Operating at equal sensitivity and specificity, the GMM detector produced 4% error while the broadband energy detector produced 32% error and the peak energy detector produced 8% error. The peak energy

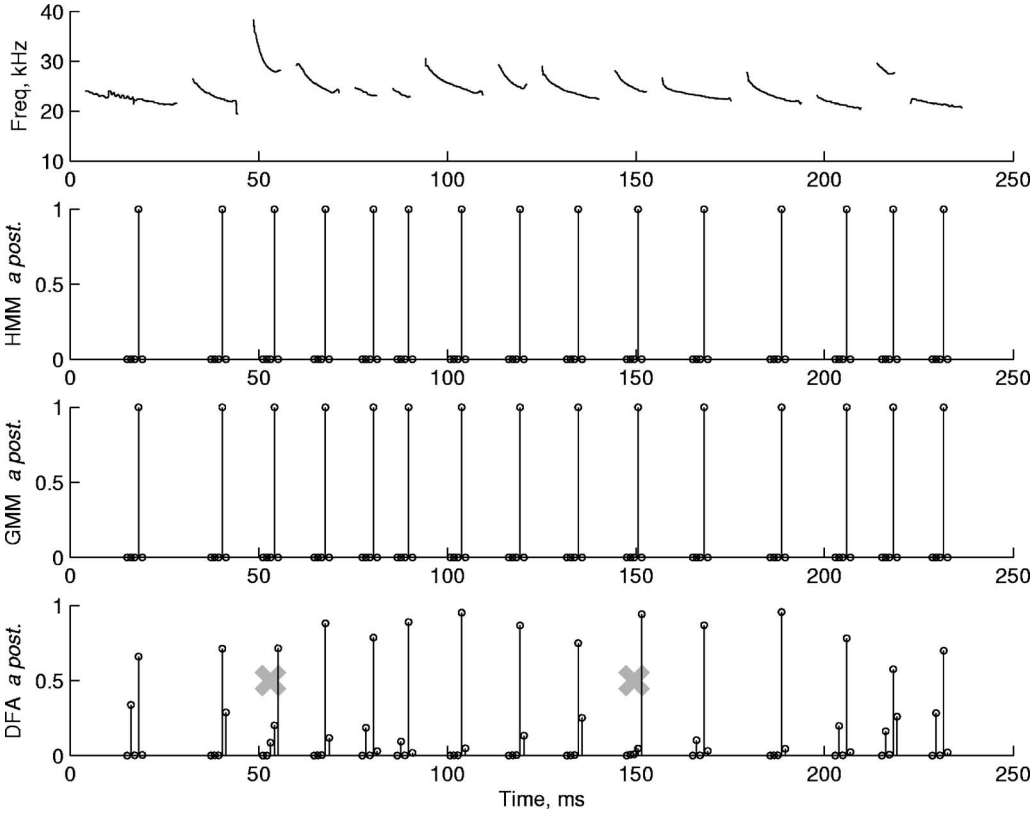


FIG. 10. Example sonograms and detector *a posteriori* probabilities for 15 randomly selected calls of *L. cinereus semotus*. Each call and corresponding groups of output probabilities from each detector are aligned vertically. The order of output probabilities for each call, left to right, is *P. bodenheimeri, M. molossus, L. borealis, L. cinereus semotus*, and *T. brasiliensis*. An "X" behind a group of probabilities indicates incorrect classification.

detector employed SMS, which significantly improved detector performance over the broadband energy baseline. The GMM detector also employed SMS and further reduced detection error compared to the peak energy detector by incorporating more information (i.e., features) from the signal and also by accounting for the characteristics of the background signal.

The differences in performance among the detectors are

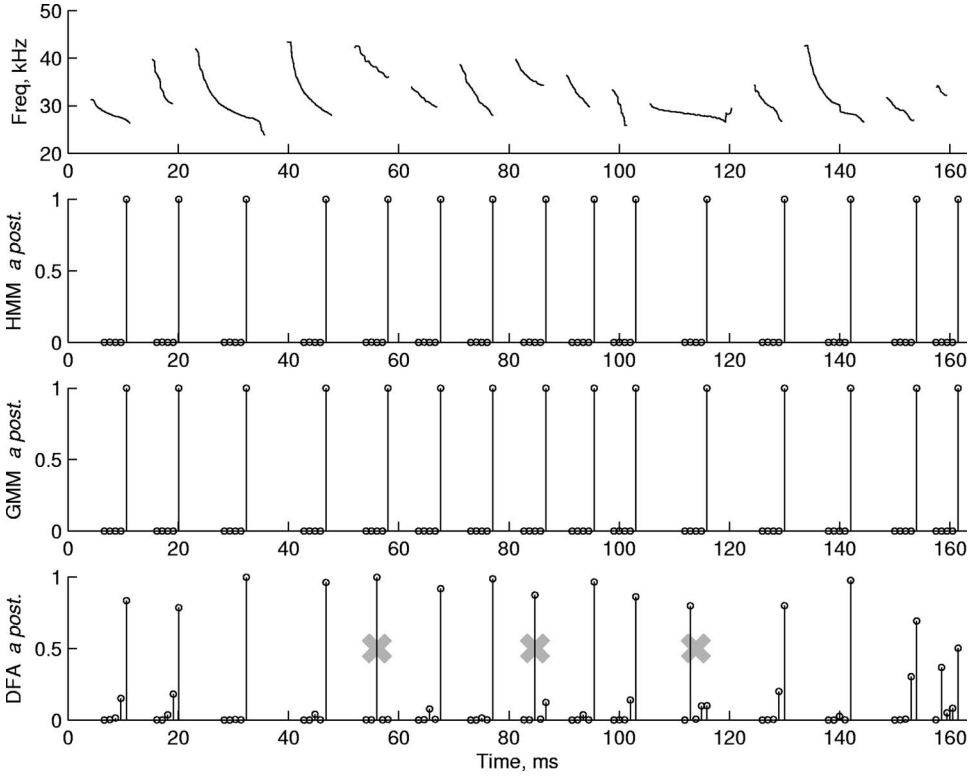


FIG. 11. Example sonograms and detector *a posteriori* probabilities for 15 randomly selected calls of *T. brasiliensis*. Each call and corresponding groups of output probabilities from each detector are aligned vertically. The order of output probabilities for each call, left to right, is *P. bodenheimeri, M. molossus, L. borealis, L. cinereus semotus*, and *T. brasiliensis*. An "X" behind a group of probabilities indicates incorrect classification.

demonstrated in Fig. 6 for a single pass of calls. The broadband energy detector only clearly registered the calls in the center of the pass, presumably when the bat was closest to the microphone. The hand-labeled calls at the beginning of the pass produced only small changes in the broadband energy detector output compared to the range of output from the background, while the calls at the end of the pass were not distinguishable from the background. The peak energy detector provided more distinction of the calls at the ends of the pass, although the peaks in the hand-labeled regions were close to those in the background regions. On the other hand, the GMM output produced significant peaks throughout the pass, relying not only on energy information but also on frequency and slope information. Calls that appeared to be "in the noise" according to the bottom panel in Fig. 6 were detected with nearly the same log likelihood as calls in the center of the pass.

The detection experiment discretized time by using 20-ms bins, which was a simpler experiment to implement compared to detecting the endpoints of hand-labeled calls. Endpoints were subjectively determined during the hand labeling of the data, and the ends of calls were particularly difficult to label due to the presence of echoes. While identifying the endpoints in a detection experiment would have been a more relevant task, the difficulty in implementing such an experiment would obfuscate the comparisons among traditional call detection methods and machine learning methods. Another step in the detection process, going from 20-ms bins to endpoints, could be added, but the current experiment using only discrete bins demonstrates sufficiently the significant improvement in call detection that is achievable using machine learning methods.

Detecting more calls from a single pass is important for several reasons. The most obvious advantage is increased detection range. Judging by the width of the range of significant detector outputs from the broadband energy detector in Fig. 6(c), the GMM detector increased detection range by nearly a factor of 2. Increased detector range means that more useful data are collected from a single recording session and also that fewer detectors are needed to cover a given range when collecting multiple simultaneous recordings, lowering equipment costs and also the amount of raw data to store and process. Another advantage involves the use of multiple calls from an individual bat. By increasing detection range, more data from an individual bat can be collected and used for individual identification or for studies involving the variety of calls made by an individual depending on factors of the surrounding environment. Furthermore, several species of echolocating bats emit low-intensity calls which are only detectable within a range of about 1 m using conventional automated algorithms. Increasing the range of detection for these species using machine learning methods would significantly increase the amount of useful data collected.

C. Classification

Two classification experiments were performed and offer complementary results for the five species under study. The resubstitution experiment produced the confusion matrices in Tables IV–VI which show about 0.3% classification error for the machine learning algorithms and about 17% error for the baseline DFA classifier. The majority of errors for the machine learning algorithms came from calls from *M. molossus* that were classified as *T. brasiliensis*, both in the Molossidae family, although there were almost no errors of *T. brasiliensis* being classified as *M. molossus*.

The DFA classifier made significantly more classification errors, ranging from 3% error for *P. bodenheimeri* to 30% error for *T. brasiliensis*, although the classification errors were not symmetrically distributed according to Table VI. The errors for calls from *L. cinereus semotus* were mainly from *T. brasiliensis* and vice versa, although *T. brasiliensis* calls were also misclassified as *M. molossus* and *L. borealis*. Calls from *P. bodenheimeri* were also misclassified most often as *L. borealis*, although very few calls were misclassified as *P. bodenheimeri*. Most of the classification errors for calls from *L. borealis* were misclassified as *M. molossus*, while errors for calls from *M. molossus* were mostly misclassified as *L. cinereus semotus*.

The cross-validation classification experiment produced the confusion matrices in Tables VII–IX which show $0.6 \pm 0.2\%$ classification error for the machine learning algorithms and $16.9 \pm 1.1\%$ error for the baseline DFA classifier. Randomly selecting half of all calls for training and the other half for testing in each of 20 trials, the resubstitution classification error for the HMM and GMM classifiers were about one standard deviation lower than those from the cross-validation experiment, owing to the reduction in training data and the novelty of the test calls. On the other hand, the resubstitution error for the DFA classifier was well within one standard deviation of the DFA cross-validation error.

The GMM and DFA classifiers are very similar in that both model the pdf of features and use Bayesian statistics to classify test data. However, the DFA uses only a single Gaussian kernel to represent the distribution of features for each species, while the GMM uses a mixture of kernels which better models the distribution of features. Furthermore, the DFA only considers a subset of the data available to the GMM. The HMM adds a temporal constraint to the GMM through the use of a state transition matrix, although the results from the current experiments did not distinguish the two models due to ceiling effects in classification accuracy.

A limitation of the current classification experiments is the fact that each species of bat was recorded from a single location, separate from all other species. Characteristics of each recording site and of each recording system could influence the feature extraction, although SMS strives to eliminate the transfer function of a recording environment (i.e., the room transfer function, microphone and amplifier transfer function, and gain settings). The confusion matrices in Tables VI and IX show that errors for the baseline classifier were rarely among species recorded with the same system. For example, a Pettersson system was used to record calls from *P. bodenheimeri* and *M. molossus*, but *P. bodenheimeri* was mostly confused with *L. borealis* while *M. molossus* was mostly confused with *L. cinereus semotus*. Both *Lasiurus* species were recorded with an Avisoft system.

Using multiple recordings from multiple sites of a single species would allow the performance of the classifiers to be characterized by training the classifiers using calls from one recording session and testing using calls from another session. Such an experiment design would also eliminate the possibility of testing and training on calls from the same individual bat, provided the different recording sessions are sufficiently separated in time and space. Unfortunately, the data collected for the current work precluded such an experiment.

D. Example plots

The example sonograms shown in Figs. 7–11 and accompanying detector outputs show several aspects of the challenge of automated detection and classification. In each of the five figures, the top panel shows 15 randomly selected sonograms from one of the species, created using the frequency feature extracted for each frame of the hand-labeled calls. The most significant characteristic of the sonograms is the variability of calls within each species. The calls from *M. molossus* in Fig. 8 are particularly varied, separated into two clusters of frequency ranges. The jaggedness of the calls in Fig. 8 is mainly due to noise in the original recording (most likely short-delayed echoes) that interferes with the estimation of power spectrum peaks. Despite such noisy sonograms, the machine learning algorithms reliably model and classify such calls.

Another aspect of the sonograms is the varied length of calls within each species. Many labeled calls are low in intensity, near the noise floor, and only the most-intense portions of such calls are visible in the hand-labeling GUI. Thus, many calls are short fragments, similar to portions of longer calls recorded from each species. Some would consider such fragmented calls useless and worth discarding, but the classification results indicate that such calls contain sufficient information to be correctly classified by the machine learning algorithms. The DFA classifier, on the other hand, appears to make the majority of its errors on fragmented calls. The baseline classification accuracy would undoubtedly increase if the fragmented calls were removed from the experiment, although at the expense of decreasing the effective range of the detector. The hand labeling performed for the current experiments strove for complete documentation of all spectrogram features which appeared to be echolocation calls, no matter how degraded the call looked in the spectrogram, and the machine learning classification results prove that all calls, even fragmented calls, contain relevant information about the bat producing those calls.

E. Machine learning limitations

The current implementations of machine learning algorithms have several limitations. The most significant limitation is the fact that the models require significant amounts of hand-labeled data to reliably train. Collecting accurate ground truth information for each call is tedious, subjective, and error prone. While the DFA classifier may be trained on a few dozen calls, the machine learning algorithms require hundreds of calls because of the large number of parameters used in the pdf models. The DFA classifier used in the classification experiment contained 20 parameters per pdf model while the GMM classifier contained 1425 parameters per pdf model.

Because the machine learning algorithms contain significantly more parameters to train than the baseline algorithms, testing and training using the machine learning algorithms is relatively slow. While real-time operation is achievable with machine learning algorithms, the current implementations required approximately ten times real time to extract features and test calls using the models on a 3.4 GHz P4 Dell Dimension 4600 desktop computer (Dell Inc., Round Rock, TX, USA). That is, 1 min of recorded data required 10 min to process (i.e., apply SMS and extract features) and use as test input for detection or classification. Training the models also took significantly longer than real time, but in most applications training would be performed off-line. The current work did not explore optimizing the settings used in the experiments, but the trade-offs between speed and accuracy of the models are discussed in the next section.

Another noteworthy aspect of the machine learning algorithms is the modeling of background information by the GMM detector. In general, the background noise may vary considerably throughout a single recording session and also among recording sessions, so trying to model "everything that is not a bat call" would prove to be an ill-conditioned problem. In practice, the noise encountered is limited to a few sources, notably electronic noise from the recording equipment and from environmental sources like rustling foliage and sympatric fauna. However, a novel stimuli that is very different from a given background model may appear to be a bat call even though the likelihood from the call model might be relatively low compared to known bat calls.

F. Settings optimization

The current experiments demonstrated the superior performance of machine learning algorithms compared to the existing algorithms employed for bat call detection and classification. However, the machine learning algorithms used several parameters that were set by hand for the current experiments: frame length, frame rate, FFT length, number of states per HMM, number of kernels per model, temporal derivative window length, and other parameters related to the training algorithms (e.g., number of training iterations and error tolerance). The parameters for frame length, frame rate, and temporal derivative window length were determined empirically in previous studies[26,27] and maintained for the current experiments. The remaining parameters were determined empirically as part of the design of the current experiments.

The choice of parameters affects the accuracy of the models and also the time required to execute the experiments. In the current work, the trade-off between speed and accuracy emphasized accuracy in order to demonstrate the performance gains achievable using machine learning methods. Lower frame rates, shorter windows, shorter FFT lengths, fewer input dimensions, and fewer Gaussian kernels

would all speed up the experiment execution. Variations in performance of the machine learning algorithms due to changes in those parameters is an interesting topic but beyond the scope of the current work.

The questions regarding the tradeoff between speed and accuracy are similar to those regarding recording source and accuracy. All of the data used in the current experiments were recorded with broadband microphones and storage, but data from heterodyne or frequency division bat detectors may be suitable for use with machine learning algorithms despite the distortion of the data from such detectors. The distorted data may retain enough discriminating information for the machine learning algorithms to learn, although accuracy would most likely suffer given the severity of the distortions typically sustained from such detectors. Furthermore, systems using zero-crossing meters may also be suitable for use with machine learning algorithms, although significant differences in bat call features between time-expanded and zero-crossing-meter outputs were previously noted.[28]

G. Implementation

To use the machine learning methods for detection and classification described in the current work, a user would need two parts: (1) a set of functions to implement the algorithms for feature extraction, noise reduction, and model training and evaluation and (2) a set of hand-labeled bat calls. We implemented the algorithms using Matlab, relying on Murphy's Full BNT toolbox for training and evaluation of the models.[24] The feature extraction and noise reduction algorithms use only simple functions, such as an FFT or logarithmic and exponential transforms, and can be implemented using a wide variety of software languages and computer platforms. To generate hand labels for training, we used our feature extraction and noise reduction algorithms to enhance a spectrogram depiction of the recordings of echolocating bats that we collected. Once the ground truths for detection and classification from the hand labels were established, the model parameters for the machine learning algorithms were estimated using the functions in the Full BNT toolbox.

V. CONCLUSIONS

Automatic speech recognition research has inspired several methods employed in the current work that have not been considered in previous bat acoustic studies: machine learning algorithms, spectral mean subtraction, and Δ and $\Delta\Delta$ features. The experiment results in the current work demonstrated that machine learning algorithms were superior to existing technology for both automated detection and classification of echolocation calls. As ASR research discovered 2 decades ago for human speech, machine learning algorithms use more information and account for variations in the data better than human-expert based methods.

Spectral mean subtraction removed two sources of variation present in the recorded data: the recording environment transfer function and harmonic noise from the recording equipment. Three recording systems were used to gather the data used in the current experiment, potentially biasing classification errors towards species recorded with the same systems. However, the confusion matrices for all classification experiments indicate that no such biasing occurred. Temporal information was embedded in each frame of features using Δ and $\Delta\Delta$ features. While duration was not explicitly used as input to the machine learning algorithms, the temporal features provided an alternative method of including shape information of the calls into the models. SMS and the functions that generate Δ and $\Delta\Delta$ features are simple yet effective algorithms which have stood the test of time in ASR research.

The machine learning experiments required little user interaction other than providing ground truth data for training. No assumptions were made regarding the frequency range of interesting calls, and no recording-specific filtering was required to reduce noise. Moreover, no calls were arbitrarily removed from the experiment because they "looked" like poor representations of the calls from a particular species. As the top panels in Figs. 7–11 show, all species produced a variety of calls of different bandwidths and durations, yet the machine learning algorithms learned to correctly classify such calls 994 times for every 1000 calls used in the cross-validation experiments. With little user interaction required to detect and classify calls, the methods employed in the current work can be used as an off-line tool to process and analyze field recordings. Real-time detection and classification of calls *in the field* are also possible once the algorithm implementations have been sufficiently optimized to reduce execution time, allowing bat researchers to spend less time managing data and more time collecting data.

While machine learning models are capable of learning the variations of features for a given training set, the algorithms are not capable of revealing *what* they learned about the variations present in the data. Echolocation calls from a single species vary because the needs of the bat vary as it interacts with its environment, and machine learning algorithms are capable of uncovering the meanings behind such variations only when the algorithms are presented with sufficient training data. However, with enough labeled data, machine learning algorithms offer powerful alternative methods for bat acoustic research.

ACKNOWLEDGMENTS

The authors would like to thank Brock Fenton for kindly providing the Avisoft and Pettersson data and also Graeme Cumming for many stimulating discussions on various aspects of the research topic. The authors would also like to acknowledge Rob Olsen for supervising the construction of the portable amplifier/anti-aliasing filter.

[1]D. R. Griffin, *Listening in the Dark: The Acoustic Orientation of Bats and Men* (Comstock, Ithaca, NY, 1958).
[2]M. B. Fenton and G. P. Bell, "Recognition of species of insectivorous bats by their echolocation calls," J. Mammal. **62**(2), 233–243 (1981).
[3]M. K. Obrist, "Flexible bat echolocation: the influence of individual, habitat and conspecifics on sonar signal design," Behav. Ecol. Sociobiol. **36**, 207–219 (1995).
[4]W. L. Gannon, R. E. Sherwin, and S. Haymond, "On the importance of articulating assumptions when conducting acoustic studies of habitat use

by bats," Wild. Soc. Bull. **31**(1), 45–61 (2003).
[5]D. Russo and G. Jones, "Identification of twenty-two bat species (Mammalia: Chiroptera) from Italy by analysis of time-expanded recordings of echolocation calls," J. Zool., London **258**(1), 91–103 (2002).
[6]M. B. Fenton, H. G. Merriam, and G. L. Holroyd, "Bats of Kootenay, Glacier, and Mount Revelstoke national parks in Canada: identification by echolocation calls, distribution, and biology," Can. J. Zool. **61**, 2503–2508 (1983).
[7]S. Parsons, A. M. Boonman, and M. K. Obrist, "Advantages and disadvantages of techniques for transforming and analyzing chiropteran echolocation calls," J. Mammal. **81**(4), 927–938 (2000).
[8]M. K. Obrist, R. Boesch, and P. F. Fluckiger, "Variability in echolocation call design of 26 Swiss bat species: consequences, limits and options for automated field identification with a synergetic pattern recognition approach," Mammalia **68**(4), 307–322 (2004).
[9]M. J. O'Farrell, B. W. Miller, and W. L. Gannon, "Qualitative identification of free-flying bats using the Anabat detector," J. Mammal. **80**(1), 11–23 (1999).
[10]C. J. Weinstein, S. S. McCandless, L. F. Mondshein, and V. W. Zue, "A system for acoustic-phonetic analysis of continuous speech," IEEE Trans. Acoust., Speech, Signal Process. **23**(1), 54–67 (1975).
[11]B.-H. Juang and S. Furui, "Automatic recognition and understanding of spoken language-a first step toward natural human-machine communication," Proc. IEEE **88**(8), 1142–1165 (2000).
[12]M. Tang, S. Seneff, and V. Zue, "Two-stage continuous speech recognition using feature-based models: a preliminary study," in *IEEE Automatic Speech Recognition and Understanding Workshop* (IEEE, St. Thomas, U.S. Virgin Islands, 2003), pp. 49–54.
[13]L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Readings in Speech Recognition*, edited by A. Waibel and K.-F. Lee (Kaufmann, San Mateo, CA, 1990), pp. 267–296.
[14]A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1989).
[15]S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust., Speech, Signal Process. **29**(2), 254–272 (1981).
[16]S. Parsons, C. W. Thorpe, and S. M. Dawson, "Echolocation calls of the long-tailed bat: A quantitative analysis of types of calls," J. Mammal. **78**(3), 964–976 (1997).
[17]S. Parsons and G. Jones, "Acoustic identification of twelve species of echolocating bat by discriminant function analysis and artificial neural networks," J. Exp. Biol. **203**(17), 2641–2656 (2000).
[18]W. M. Masters, S. C. Jacobs, and J. A. Simmons, "The structure of echolocation sounds used by the big brown bat *Eptesicus fuscus*: Some consequences for echo processing," J. Acoust. Soc. Am. **89**, 1402–1413 (1991).
[19]B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am. **55**, 1304–1312 (1974).
[20]R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. (Wiley, New York, 2001).
[21]C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford U. P., New York, 1995).
[22]R. F. Lance, B. Bollich, C. L. Callahan, and P. L. Leberg, "Surveying forest-bat communities with Anabat detectors," in *Bats and Forests Symposium*, edited by R. M. R. Barclay and R. M. Brigham (Research Branch, British Columbia Minister of Forestry, Victoria, BC, 1996), pp. 175–184.
[23]R. A. Krusic and C. D. Neefus, "Habitat associations of bat species in the White Mountain National Forest," in *Bats and Forests Symposium*, edited by R. M. R. Barclay and R. M. Brigham (Research Branch, British Columbia Minister of Forestry, Victoria, BC, 1996), pp. 185–198.
[24]K. Murphy, Bayes Network Toolbox for Matlab, 2005, URL: http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html, 21 October 2005.
[25]Y. T. Chan, J. M. M. Lavoie, and J. B. Plan, "A parameter estimation approach to estimation of frequencies of sinusoids," IEEE Trans. Acoust., Speech, Signal Process. **29**(2), 214–219 (1981).
[26]M. D. Skowronski and J. G. Harris, "Statistical automatic species identification of microchiroptera from echolocation calls: lessons learned from human automatic speech recognition," J. Acoust. Soc. Am. **116**, 2639 (2004).
[27]M. D. Skowronski and J. G. Harris, "Automatic detection of microchiroptera echolocation calls from field recordings using machine learning algorithms," J. Acoust. Soc. Am. **117**, 2552 (2005).
[28]M. B. Fenton, S. Bouchard, M. J. Vonhof, and J. Zigouris, "Time-expansion and zero-crossing period meter systems present significantly different views of echolocation calls of bats," J. Mammal. **82**(3), 721–727 (2001).

Effects of nonlinear propagation, cavitation, and boiling in lesion formation by high intensity focused ultrasound in a gel phantom

Vera A. Khokhlova[a)]
Department of Acoustics, Faculty of Physics, Moscow State University, Moscow, 119992, Russia and Center for Industrial and Medical Ultrasound, Applied Physics Laboratory, University of Washington, Seattle, Washington 98105

Michael R. Bailey, Justin A. Reed, Bryan W. Cunitz, Peter J. Kaczkowski, and Lawrence A. Crum
Center for Industrial and Medical Ultrasound, Applied Physics Laboratory, University of Washington, Seattle, Washington 98105

(Received 8 January 2005; revised 5 December 2005; accepted 5 December 2005)

The importance of nonlinear acoustic wave propagation and ultrasound-induced cavitation in the acceleration of thermal lesion production by high intensity focused ultrasound was investigated experimentally and theoretically in a transparent protein-containing gel. A numerical model that accounted for nonlinear acoustic propagation was used to simulate experimental conditions. Various exposure regimes with equal total ultrasound energy but variable peak acoustic pressure were studied for single lesions and lesion stripes obtained by moving the transducer. Static overpressure was applied to suppress cavitation. Strong enhancement of lesion production was observed for high amplitude waves and was supported by modeling. Through overpressure experiments it was shown that both nonlinear propagation and cavitation mechanisms participate in accelerating lesion inception and growth. Using B-mode ultrasound, cavitation was observed at normal ambient pressure as weakly enhanced echogenicity in the focal region, but was not detected with overpressure. Formation of tadpole-shaped lesions, shifted toward the transducer, was always observed to be due to boiling. Boiling bubbles were visible in the gel and were evident as strongly echogenic regions in B-mode images. These experiments indicate that nonlinear propagation and cavitation accelerate heating, but no lesion displacement or distortion was observed in the absence of boiling. [DOI: 10.1121/1.2161440]

PACS number(s): 43.80.Gx, 43.25.Yw, 43.25.Cb [CCC] Pages: 1834–1848

I. INTRODUCTION

High intensity focused ultrasound (HIFU) is an emerging modern therapy in which ultrasound energy is locally absorbed in a millimeter-size focal region to induce tissue necrosis or cauterize bleeds without damaging intervening tissues. Current experimental and clinical therapies employ various noninvasive external (for treating brain, abdominal, bone, and ocular tumors, and for acoustic hemostasis), endoscopic and other intracavitary (for treating prostate and uterine tumors) methods as well as minimally invasive intravascular and intraoperative percutaneous techniques.[1] The main technical limitations to widespread clinical use of HIFU have been the long time required to treat tumors of typically several cubic centimeter volume,[2] the lack of developed real-time imaging methods for targeting and monitoring HIFU,[3] and the large power demands that complicate portability for trauma applications such as battlefield hemostasis.[4] Most clinical results to date have been obtained in China where over 10 000 cancer patients have been treated in the past 5 years.[5] Many of these patients were treated with a HIFU device that operates at very high acoustic intensity (more than 10 kW/cm^2) combined with fast mechanical movement of the focus (several mm/s) over the tumor volume. The treatment site generally appears as a bright (echogenic) region in a B-mode ultrasound image, which is used for guiding the therapy.[6]

The use of B-mode ultrasound image-guidance while working with very high acoustic intensities and rapidly sweeping the focus was a successful innovation and appeared to be a practical solution to the first two problems of shortening treatment duration and monitoring the treated site in real time. Ultrasound image-guided HIFU devices have been developed for hemostasis and other applications.[7] More sensitive ultrasound-based techniques are under development and MRI methods are also employed for targeting and monitoring the treatment.[8–11] However, despite the existence of clinical devices, several important physical issues of nonlinear ultrasound-tissue interaction at this very high acoustic level remain poorly understood. These include the physical mechanisms responsible for the observed increase in volume and acceleration of tissue necrosis, and if this enhancement can be optimized by modifying the acoustic waveform to shorten treatment times or to reduce the energy requirements for portable trauma devices; the mechanism of lesion shape

[a)]Author to whom correspondence should be addressed. Electronic mail: vera@acs366.phys.msu.ru

distortion from a symmetric "cigar" shape to a "tadpole" shape translated from the focus towards the transducer; accurate prediction of the final tissue damage; the origin of the echogenic region on a B-mode image and B-mode sensitivity in monitoring the treatment. This present work is intended to provide deeper insight towards answering these questions.

Two nonlinear effects are known to be involved in lesion production by HIFU: nonlinear acoustic wave propagation,[4,12,13] and cavitation, i.e., ultrasound-induced oscillations of microbubbles grown from gaseous nuclei in tissue.[14–16] The relative role of these two basic nonlinear phenomena is not yet well understood because models that fully represent heating response (and ultimately, biological response) to HIFU exposure in tissue do not exist. Both effects are expected to act simultaneously at high acoustic pressure and both are responsible for acceleration of lesion inception, overall growth, and distortion of the final lesion shape.[17–19] The effect of acoustic nonlinearity results in generation of higher harmonics during propagation to the focus of the transducer, with shock formation under certain conditions for which acoustic energy is more effectively absorbed and converted to heat in the tissue.[20,21] The presence of cavitation also leads to enhanced heating of tissue through the absorption of waves scattered or emitted by the bubbles (particularly as energy is often converted to higher, more readily absorbed frequencies), diffusion of heat from the hot compressed gas of the bubble interior, and viscous damping of bubble oscillations by the tissue or body fluids. The relative importance of these three cavitation mediated mechanisms is currently under investigation.[16,22]

In addition to cavitation, bubbles in tissue may form from boiling. Boiling, that is, the growth of vapor bubbles occurs due to HIFU-induced temperature rise. In this sense, boiling is to be distinguished from cavitation which occurs due to HIFU-induced pressure oscillations. Cavitation bubbles are expected to be small (resonant size is of the order of microns for HIFU frequencies) and predominantly gaseous, as the result of rectified diffusion.[15,23] They might slowly grow by coalescence or enhanced outgassing from the fluid presumed to be saturated with gas due to elevated temperature, but they can also fission as a result of violent collapses called inertial cavitation. On the other hand, vapor bubbles created by boiling are not directly caused by pressure oscillations and thus can grow rapidly to a large size (on the order of millimeters). This growth is particularly explosive if super heating occurs as a result of rapid temperature increase. Several large boiling bubbles may scatter and reflect ultrasound much more effectively than a cavitating microbubble cloud. Strong reflections result in shielding the focus from HIFU energy and increased prefocal heating. Boiling bubbles, more so than cavitation bubbles, may therefore contribute to distortion of the lesion from a cigar shape into a tadpole shape and growth of the lesion toward the transducer.[24] Scattering from bubbles also produces a region of increased echogenicity on a B-mode image but it has been difficult to distinguish between cavitation bubbles and boiling bubbles while monitoring HIFU.[3,25,26]

The goal of this work was to investigate experimentally and numerically the relative importance of acoustic nonlinearity and bubble dynamics (cavitation and boiling) in enhancement of heat deposition by HIFU, B-mode ultrasound imaging of treatment, and lesion dynamics. A transparent protein-containing gel phantom that turns optically diffusive as proteins denature was used to visualize heated volumes.[27] Here we refer to these whitish diffusive regions as lesions by analogy to bioeffects observed in tissue. Although it is not possible to directly transfer the results of observations in gel to real tissues, because some properties of gel differ from tissue, the gel permits direct optical visualization of lesion growth in real time (not possible in tissue) correlated with simultaneous observations using B-mode ultrasound (possible in tissue). Furthermore, numerical modeling of HIFU in gel indicates that the changes due to nonlinear propagation are relatively greater than in tissue thus facilitating the observation of physical processes involved in lesion dynamics. The appearance of large bubbles, for example, due to boiling, can clearly be observed in the gel, even inside the diffusive lesion under backlighting conditions.

Experiments were conducted in two different physical arrangements. In the first set of experiments, in addition to single point lesions it was possible to produce lesion stripes, but only under standard atmospheric pressure. In the second set of experiments, elevated static pressure was used to suppress cavitation and boiling, but only single lesions were produced due to size constraints of the hyperbaric chamber. Different nonlinear regimes of heating were studied based on the same time-average acoustic power of the transducer but different pulse power (peak pressure). Metrics of lesion inception time, lesion growth rate, the appearance of bubbles, and morphological change of the lesion from symmetric cigar to distorted tadpole shape were used to characterize different regimes and stages of lesion dynamics.

II. EXPERIMENTAL METHODS

A. Tissue-mimicking phantom

An optically transparent polyacrylamide gel containing Bovine Serum albumin (BSA) was used in experiments as a tissue-mimicking phantom. The transparent gel turns translucent at high temperatures near 60 °C due to the presence of BSA proteins that become optically diffusive when thermally denatured. Table I lists the ingredients and proportions required to create a 200 ml gel with 7% concentration of BSA. This recipe was recently developed for use as a thermal in-

TABLE I. Chemical composition of 200 ml polyacrylamide gel with 7% BSA concentration.

	ml	Percent
Total volume	200.000	
Distilled water	143.220	0.7161
1 M TRIS	20.000	0.1000
40% Acrylamide	35.000	0.1750
10% APS	1.680	0.0084
TEMED	0.100	0.0005
	Grams	
BSA	14	7.000

dicating phantom for HIFU,[27] and it is included here for completeness. The gel phantoms were prepared by first mixing the BSA protein (Sigma-Aldrich) in distilled water degassed by pinhole degassing. The solution was gently stirred to mix the BSA powder, which likely reintroduced some gas. A solution of acrylamide then was added to the mixture, followed by a 1M TRIS buffer *p*H 8 (trizma hydrochloride and trizma base, Sigma-Aldrich), and the APS (ammonium persulfate, Sigma-Aldrich). The entire solution was placed in a vacuum chamber and held under vacuum of 700 mm Hg strength for over 1 hour for additional degassing. The final polymerization agent was degassed in a vacuum chamber and added to the remaining solution which was then transferred into cylindrical or cubical molds, and allowed to polymerize under vacuum. For comparison, some gels were made without degassing the ingredients. The gels have a useful lifetime of several weeks and these experiments were conducted the day of manufacture or the gels were vacuum-sealed in plastic bags, stored at 5 °C, and used the next day.

It has been shown that the acoustical and thermal properties of the gel are similar to those of tissue, although the acoustic attenuation in gel with 7% BSA is about one-third of the value in tissue.[27] Higher protein concentration would result in stronger absorption, but less transparency of the gel which becomes cloudy. Using a higher percentage of acrylamide makes the gel stiffer but does not appreciably increase attenuation; furthermore, the exothermic polymerization reaction becomes too hot to prevent denaturation of the BSA during gel preparation.

B. Experimental arrangement for lesion stripes

For the first set of experiments, performed under atmospheric pressure, the experimental arrangement presented in Fig. 1 allowed production of both static lesions and lesion stripes obtained by linearly translating the ultrasound transducer. A cubical gel sample of 5.5 cm on a side was mounted in a large water tank at room temperature (22 °C). The water in the tank was degassed for 2 hours prior to the measurements, which resulted in an air saturation level of less than 25% measured by a dissolved oxygen meter (YSI, Yellow Springs, Ohio). The sample was held in position by an acrylic box with open sides facing and opposite the transducer to prevent reflection from the acrylic. A CCD camera was used to film lesion formation with fiber optic lighting from above, and to record the image to VHS video tape. An LED placed in the field of view of the CCD indicated when the HIFU was on. Videotapes were later digitized by an ATI video capture board and software. Observation time of different stages of lesion development was determined by counting video frames (30 fps).

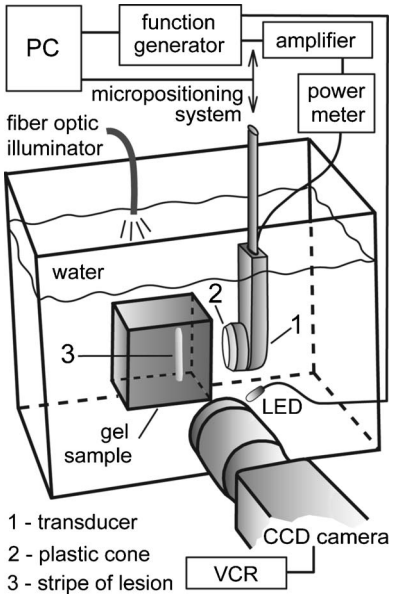


FIG. 1. Experimental arrangement for video imaging of HIFU-induced single lesions and lesion stripes in gel phantom. The transducer of 3.5 MHz frequency, 33 mm aperture, and 35 mm focal length was translated vertically at a constant velocity while HIFU was on.

The transducer (SU-107 Sonic Concepts, Woodinville, WA) had a 3.5 MHz single element with 33 mm aperture and 35 mm radius of curvature. A truncated hollow acrylic cone larger than the acoustic beam width of the HIFU was used to aid in alignment and to hold a thin polyethylene film to suppress the development of streaming in the water path between the transducer and the sample. The transducer was affixed to a computer controlled 3-axis positioning system and a LABView (National Instruments, Austin, TX) program was written to control the experiment, including setting the timing and communicating the desired signal amplitude to the HP 33120 function generator via GPIB. The RF signal generated by the function generator was amplified by a linear 150 W RF amplifier (ENI A150). Forward RF power was monitored by a power meter (Sonic Concepts, Woodinville, WA) that measured voltage and current, and displayed true time-averaged electric power into the transducer. For comparison to the results of modeling the transducer was calibrated by a radiation force balance method to measure acoustic power output versus electrical power input at the levels of ultrasound used in the experiments. The efficiency was found to be $84\% \pm 4\%$. The ultrasound beam pattern was also measured in water using a calibrated needle hydrophone (SEA model GL-0150-1A with 150 μm active area, Soquel, CA) at low power to ensure a good match of the experimental pressure field geometry with the predictions of the theoretical model.

For static single lesions, preliminary experiments were performed in continuous wave (cw) regime with gradually increased transducer power to observe the changes in lesion inception and further growth. A baseline power value was empirically selected such that a lesion appeared in 4 seconds, which is a typical exposure for HIFU applications. For higher power levels, small steps in power increase were used to observe the appearance of bubbles and qualitative changes in lesion dynamics within the 4 second exposure. Further measurements were conducted with variation of pulse acoustic power combined either with the same exposure time (4 seconds) and different duty cycle, or with different exposure times with cw exposure, such that equal total acoustic energy was transmitted into the gel. When the duty cycle was less than 100%, the value of pulse acoustic power was obtained by dividing the cw power level of the transducer by the chosen duty factor to keep the same time-averaged value of

acoustic power. The pulse repetition frequency for the duty cycle regime was 1 kHz, and the duty cycle varied from 100% to 50%, with corresponding doubling of the pulse power level. The idea of exploiting nonlinearity to enhance heating with constant energy input (by varying duty cycle) has been presented earlier.[12,28]

When producing stripes of lesions, the LABView program was used to control the speed with which the transducer was translated and to turn on the HIFU only when the transducer was not accelerating. Signal generation was initiated with some delay to permit the transducer to reach a steady speed. To maintain a regime of equal acoustic energy radiated into the gel, the stripes were produced with variation of duty cycle (100%–6.25% with 1 kHz repetition rate) combined either with the same transducer velocity (0.5 mm/s) and different pulse acoustic power (15–240 W), or by varying transducer velocity (0.5–6 mm/s) and preserving constant acoustic power. The power regimes were chosen based on the results of the experiments with single lesions so as to produce a visible small trace of lesion in cw regime at the lowest level of power (least nonlinear acoustic regime) at the slowest velocity of the transducer. Then the pulse power was increased incrementally in proportion to the decrease of the duty cycle or transducer velocity to maintain the same time-average acoustic power. Transducer velocity was chosen within the range of the ones typically used in clinical HIFU.[5,6] Both degassed and nondegassed gel samples were compared to evaluate the utility of degassing in gel preparation.

C. Experimental arrangement for overpressure

The goal of the second set of experiments was to apply static overpressure to isolate the effect of nonlinear propagation from the effect of bubbles in enhancement of heat deposition. Overpressure (elevated hydrostatic pressure) dissolves existing bubbles, raises the boiling temperature, and restricts the activity of any HIFU induced microbubbles or boiling bubbles that may form in gel. The experimental arrangement is shown in Fig. 2. An aluminum pressure chamber was built using a 10 by 12.5 by 15 cm aluminum block with two cross-

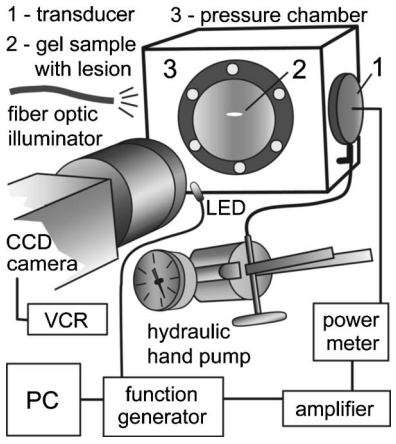


FIG. 2. Experimental arrangement for video imaging of HIFU-induced single lesions in gel under elevated static pressure. The transducer was of 2 MHz frequency, 40 mm aperture, and 44 mm focal length.

ing bore holes of 5 cm and 6.2 cm diameter made on perpendicular sides of the chamber. The chamber can withstand an overpressure of more than 120 bars. Cylindrical gel samples were 6 cm in diameter by 6 cm long. The gel sample sat in the larger diameter horizontal bore hole with degassed water added to fill the remaining volume of the chamber. An oil-backed, piezoceramic, axially symmetric spherically shaped HIFU transducer of 2 MHz frequency, 40 mm aperture, and 44 mm radius of curvature was mounted on the right-hand side of the chamber closing one end of the other bore hole. Opposite the HIFU transducer, on the left-hand side, was a 2.5 cm thick transparent acrylic window. Two more acrylic windows of the same thickness capped both ends of the larger bore hole on the front and back sides of the chamber. A fiber optic light was used to illuminate the sample either from the side through the left-hand acrylic window or from the back through a diffusive sheet of white paper placed close to the window. Lesion formation was filmed by a CCD camera through the front window of the chamber. In this set of experiments, the A150 amplifier was replaced by an AP400B model because the transducer efficiency ($39\% \pm 4\%$) and frequency were lower than that of the SU-107 transducer. The transducer was calibrated as before, using combined radiation force balance, field mapping, and modeling tools; electric power was monitored during experiments using the power meter.

The hydraulic hand pump (Ralston Instruments, Chagrin Falls, Ohio) was used to apply pressure to the water-filled interior of the closed pressure chamber with the gel sample inside. A fluid conduit connected the water to the oil that filled small high pressure transducer housing and insulated the backside of the transducer element. A free piston inside the conduit prevented mixing of the two fluids while permitting equalization of pressure between oil and water, and thus across the brittle transducer element. An electrical network matched the transducer to 50 Ohms, but no acoustic matching layer was developed for the transducer face. Net electrical power to the transducer was measured and was the same with and without overpressure.

Modifications were made to the acrylic caps of the chamber (not shown in Fig. 2) to make additional measurements with an ATL HDI-1000 ultrasound system (Philips, Bothell, WA) which was used to record B-mode images to SVHS video tape. The C4-2 broadband diagnostic imaging probe with 3 MHz central frequency and 2–4 MHz bandwidth was affixed to the chamber and coupled by commercial ultrasound coupling gel at the rear acrylic window to monitor the treatment and to correlate B-mode images with visual observations. The alignment of the imaging focus and the HIFU focus was performed by imaging the opposite chamber window. In order not to saturate the B-mode image with HIFU, the HIFU was synchronized to the imager frame rate (32 Hz), run at 50%–72% duty cycle (except under cw), and phased so that the interference was relegated to the edges of the image.

To assess the presence of cavitation for the experiments at ambient static pressure additional cavitation measurements were performed outside the pressure chamber for degassed gel samples. A passive cavitation detection (PCD) system

consisted of a 20 MHz focused transducer (Staveley Sensors Inc., East Hartford, CT) aligned confocally and with its axis perpendicular to HIFU, a 15 MHz high-pass filter (Allen Avionics Inc., Mineola, NY), a preamplifier (Panametrics-NDT Inc., Waltham, MA), and a Lecroy LT344 oscilloscope (Chestnut Ridge, NY). A 200 cycle burst was sent from the HIFU source at increasing power levels and the resulting PCD traces were recorded. The onset of cavitation was correlated with a dramatic change in PCD signal level. Using this system, it was found that the cavitation threshold of the degassed, 7% BSA-polyacrylamide gel varied over the range from 2 to 16 W acoustic power of the HIFU source, depending on how carefully the ingredients of the gel were degassed. This power range corresponds to acoustic peak negative pressure of 2–6 MPa in the gel at the focus, obtained from the results of nonlinear modeling.

All the experiments in overpressure chamber (at ambient pressure) were always performed well above the cavitation threshold and consequently cavitation was presumed to occur for all atmospheric pressure exposures. The peak acoustic power varied from 25 to 75 W and from 100% to 50% duty cycles, with a 30 s exposure that was empirically determined to provide observations of a qualitatively varied range of lesion development regimes, with and without overpressure. For overpressure experiments, samples were pressurized and maintained at pressure for 5 minutes prior to HIFU exposure to allow ample time for ambient bubbles to dissolve. In most of these experiments the level of overpressure was chosen to exceed the peak negative pressure at the HIFU focus to completely suppress cavitation.

III. THEORY AND NUMERICAL MODEL

The experiments on single static lesion formation were simulated numerically with an assumption of either linear or nonlinear ultrasound propagation.

A. Linear and nonlinear acoustic field

The HIFU field was modeled using the KZK-type nonlinear parabolic equation, generalized for the frequency dependent absorption properties of the following propagation medium:

$$\frac{\partial}{\partial\tau}\left(\frac{\partial p}{\partial z}-\frac{\beta}{\rho_0 c_0^3}p\frac{\partial p}{\partial\tau}-L_{\text{abs}}(p)\right)=\frac{c_0}{2}\Delta_\perp p. \tag{1}$$

Here p is the acoustic pressure, z is the propagation coordinate along the axis of the beam, $\tau=t-z/c_0$ is the retarded time, c_0 is the sound speed, ρ_0 is the ambient density of medium, β is the coefficient of nonlinearity, $\Delta_\perp=\partial^2/\partial r^2+r^{-1}\partial/\partial r$ is the Laplacian with respect to the transverse coordinate r, L_{abs} is the linear operator that accounts for absorption and dispersion properties of the medium.

The propagation path for ultrasound was through a two-layer medium, first in water and then in the gel sample. For simulations in water, the thermoviscous absorption was included as

$$L_{\text{abs}}=\frac{b}{2c_0^3\rho_0}\frac{\partial^2 p}{\partial\tau^2}, \tag{2}$$

where b is the dissipative parameter of water. For simulations in gel, the operator L_{abs} accounted for the power law of ultrasound absorption measured in the gel[27]

$$\alpha(f)=\alpha_0(f/f_0)^\eta \tag{3}$$

and variation of the sound speed with frequency calculated using the local dispersion relations[29] as

$$\frac{c(f)-c_0}{c_0}=\frac{c_0\alpha_0}{\pi^2(\eta-1)f_0}\begin{cases}[(f/f_0)^{\eta-1}-1], & \eta\neq 1,\\ \ln(f/f_0), & \eta=1.\end{cases} \tag{4}$$

Here α_0 is the absorption coefficient and $c_0=c(f_0)$ is chosen as the ambient sound speed at the fundamental frequency f_0. Equation (1) was solved numerically in the frequency domain using a previously developed finite difference algorithm.[21,30] The acoustic pressure waveform was represented as a Fourier series expansion; a set of nonlinear coupled differential equations for the amplitudes of harmonics was derived and integrated numerically using the method of fractional steps with an operator-splitting procedure.

Simulations were performed with and without acoustic nonlinearity in order to predict the importance of nonlinear propagation effects for particular experimental conditions. Spatial distributions of the amplitudes and the intensities I_n of the harmonics nf_0, and the total intensity of the wave $I(z,r)=\sum_{n=1}^{\infty}I_n(z,r)$ were calculated. Acoustic waveforms were reconstructed at various distances from the transducer. Heat deposition patterns due to the absorption of ultrasound

$$q_v(z,r)=2\sum_{n=1}^{\infty}\alpha(nf_0)I_n(z,r) \tag{5}$$

were obtained for further simulations of the temperature rise in the gel.

The values of the physical constants used for the modeling were $\rho_0=1000$ kg/m^3, $c_0=1486$ m/s, $\beta=3.5$, $b=4.33\times10^{-3}$ kg s^{-1} m^{-1} for water; and $\rho_0=1044$ kg/m^3, $c_0=1544$ m/s, $\beta=4.0$, $\alpha_0=1.6$ m^{-1} at 1 MHz, $\eta=1$, for the gel.[27] No changes in the acoustic parameters of gel due to HIFU heating were considered in the simulations. The linear case was modeled by choosing $\beta=0$.

B. Temperature field

Temperature rise in the gel phantom was modeled using a heat transfer equation

$$\frac{\partial T}{\partial t}=k\Delta T+\frac{q_v}{c_v}. \tag{6}$$

Here T is the temperature in the gel, c_v is the heat capacity per unit volume, k is the temperature conductivity, q_v is the distribution of thermal sources calculated from Eq. (5). Equation (6) was integrated numerically using a finite difference scheme.[21] Thermal properties of the gel in simulations were $c_v=5.3\times10^6$ J m^{-3} °C^{-1} and $k=1.3\times10^{-7}$ m^2/c.[27]

A thermal dose required to produce a lesion in tissue has been defined[31] but an analogous thermal dose for the BSA

gel is not well established. Preliminary experiments showed that gels turned optically diffusive in about 2 s at 58 °C and in less than 0.1 s at 65 °C.[27] These values are similar to those observed in tissue. Since the experiments and modeling were performed for rapid heating of gel with lesion inception time of the order of seconds, it was appropriate to choose a temperature rather than a time integrated thermal dose as a threshold of lesion inception. For calculations therefore we chose to model lesion formation by selecting a threshold temperature of 65 °C, at which the gel is considered to be instantly denatured. The lesion boundary defined in this way will be only slightly smaller than that which would be obtained by time integration.

IV. RESULTS AND DISCUSSION

A. Single lesions

The acoustic power threshold for formation of visible bubbles was found experimentally for a 4 second exposure by producing a sequence of lesions in degassed gel with gradually increasing HIFU power from 10 W to 20 W in the continuous wave (cw) regime. The lesions grew symmetrically until the end of the exposure for acoustic powers less than 16.5 W, at which formation of a big millimeter size bubble was observed (the threshold was sharp and repeatable). Figure 3 shows lesions observed in the gel at 2.5 s and 4 s for three acoustic powers close to this threshold of 16.5 W. It is seen that a very small increase in power yielded a sudden appearance of visible bubbles and corresponding significant change in lesion shape and size. The bubbles are far easier to perceive in a movie than in still images taken from the movie, but here we must present still images which were selected to best depict observations. Cigar-shaped symmetric lesions formed below a 16.5 W power threshold [Fig. 3(a)] and tadpole-shaped lesions with visible bubbles formed above the power threshold [Figs. 3(b) and 3(c)]. A large bubble was observed at the very end of the exposure at the edge of the lesion that faced the transducer for 16.5 W power [Fig. 3(b)]. Several bubbles developed in the lesion starting from 2.5 seconds for 17 W power [Fig. 3(c)]. Bubbles appeared first in the center of the lesion along the transducer axis and quickly grew towards the transducer, causing the distortion of lesion size, position, and symmetry. Significant change in lesion inception time was also observed within this small change of acoustic power, 1.8 s for 16 W, 1.5 s for 16.5 W, and 1.1 s for 17 W acoustic power.

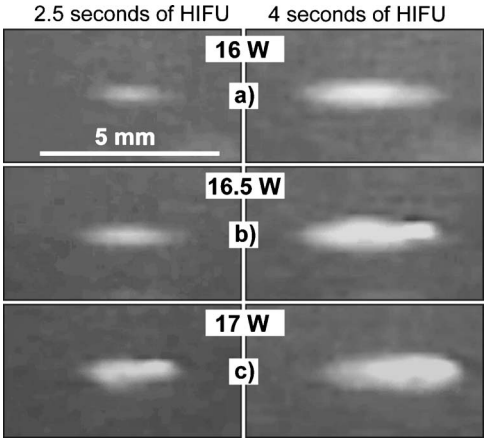


FIG. 3. Single lesions formed in degassed gel with 16 W (a), 16.5 W (b), and 17 W (c) acoustic power of the HIFU source after 2.5 s (left-hand set) and 4 s (right-hand set) exposure. The HIFU transducer was on the right-hand side. Slight increases in acoustic power cause accelerated lesion growth, then boiling, and then lesion distortion and migration toward the transducer.

Rapid changes in lesion dynamics with a small increase of the source power suggested that nonlinear effects were involved at this level of HIFU. Further experiments were performed at "high" and "low" acoustic peak pressures with the same exposure duration of 4 s, the same time-average acoustic power of 16 W, but varying the pulse power and duty cycle appropriately, either in the cw regime or with 50% duty cycle at 1 kHz repetition rate. Substantial increase of the final lesion size, almost immediate (at 0.3 s) appearance of bubble activity, and translation of the lesion toward the transducer were produced by the high-amplitude waves compared to the smooth symmetric growth of the lesion produced by low-amplitude waves. Nondegassed gel samples were also treated under the same HIFU protocol without observable difference in lesion inception time and bubble appearance as compared to the degassed samples. We expected that if cavitation played a strong role, the lesion formation would proceed differently between degassed and gas saturated gels because of rectified diffusion. However, the dynamics of lesion growth was not strongly affected by the gas content dissolved in gel, unless the HIFU hit small, but visible bubbles remaining in nondegassed samples (these voids were absent in degassed gels), which resulted in some irregular distortion of the lesion. We have proposed[4] and will elaborate here that sudden formation of visible bubbles in the lesion was due to boiling and that the fast change in lesion dynamics with a small increase of the HIFU power level was due to the formation of shocks and corresponding enhanced heating.

Similar "bubbly" structure of tadpole-shaped lesions has previously been observed visually in real tissues as cavities in the middle of the focal area and attributed to bubbles, either due to ultrasound-induced cavitation or to boiling.[32] Figure 4 shows our observations of bubbly lesions in excised bovine liver. So-called "popcorn" claps, similar to boiling noise and presumed to be due to sudden phase change of superheated fluid, have been observed during tissue treatment above some intensity level of HIFU.[33] However, the relative importance of these two bubble mechanisms under different HIFU conditions is not well understood.

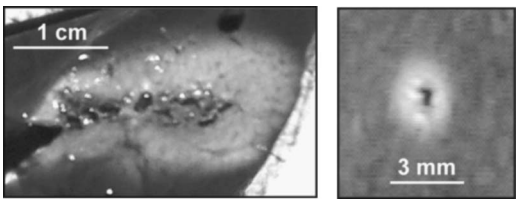


FIG. 4. Images of HIFU-induced overheated lesions in excised degassed bovine liver. Both an axial tadpole section (left-hand side) and a transverse section (right-hand side) of the lesions show vaporized cavities along the HIFU axis.

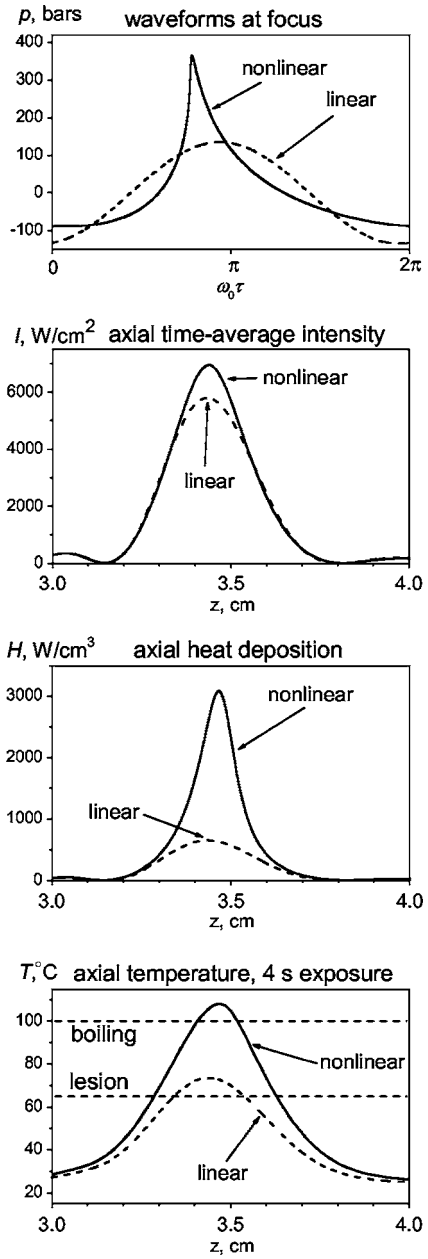


FIG. 5. Numerical results of modeling HIFU in gel with (solid lines) and without (dash lines) accounting for the effects of nonlinear propagation, axial waveforms at the focus; axial time-average intensity, heat deposition, and temperature rise after 4 s exposure. Acoustic power of the HIFU source in simulations was 16.5 W which corresponded to spatial peak intensity in gel of 5800 W/cm^2 calculated linearly. Accounting for nonlinear propagation predicts a shocked waveform, higher intensity and heat deposition, as well as boiling temperatures at focus as compared with the linear modeling.

To support the hypothesis of the effects of acoustic nonlinearity on enhanced heating and possible boiling of the gel, numerical simulations of gel heating were performed with and without nonlinear propagation effects for the conditions used in the experiments. Shown in Fig. 5 are the results of modeling of the acoustic and temperature field in the gel for an acoustic power of 16.5 W, the case in which a large bubble burst into the image at the end of the exposure [Fig. 3(b)]. The results of simulations show that at this power level of HIFU acoustic nonlinearity results in formation of a shock front in the acoustic waveform, weak enhancement of the mean intensity, and a more than fourfold increase in heating at the focus of the HIFU source. Even though linear modeling yields axial temperatures in excess of 65 °C for 4 seconds of exposure, it fails to predict boiling (temperature rise to 100 °C or more) and overestimates the lesion inception time compared to that observed in the experiments. However, nonlinear modeling produces predictions of the onset of boiling and lesion inception time that closely agree with experimental results (within about 10%). Note, that whereas the peak heat deposition on the focal axis is increased fourfold by nonlinearity, the resulting peak temperature is increased by less than 50%. The nonlinearly enhanced peak of heating is spatially much narrower than the width of the linear HIFU acoustic beam; therefore much faster temperature rise is expected at the beginning of the exposure, but rapid diffusion of the steep temperature gradient smoothes the narrow nonlinear peak more quickly than in the linear case. HIFU intensities required for rapid lesion distortion and boiling observed in gels correlated well with numerical results for the onset of shock formation close to the focus, which gives additional support to the importance of the nonlinear propagation effect in enhanced heating.

The experiments with single lesions showed that for nonlinear regimes with equal time-average acoustic power, those with higher peak pressure resulted in faster heating, lesion inception, and appearance of large bubbles. Based on the results of this section which provide an excellent correlation between experimental observations and numerical simulations of nonlinear acoustics without cavitation effects, we attributed this sudden appearance of large bubbles to boiling, and distortion of lesion shape to reflections from these vapor bubbles.

B. Lesion stripes

Nonlinearly enhanced heating might be used to accelerate treatment of a tissue volume, such as a tumor which is typically much larger than a single focal lesion. One approach to treating a volume is formation of discrete single lesions with sufficient time between lesions for cooling; this controlled approach avoids boiling and distortion of lesion shape to produce predictable, symmetric lesions.[24] Moderate HIFU intensities of 1000–2000 W/cm^2 are typically applied. A second approach involves constant mechanical movement of the transducer with a velocity between 0.5 and 4 mm/s, a typical range for HIFU clinical studies.[5] Much higher intensities are typically used to form lesion stripes in this approach and therefore nonlinear effects are expected to be more pronounced and thus benefit acceleration of the treatment. In addition to faster heating, formation of bubbles in the HIFU focal zone results in strong echogenicity that can serve to monitor the treatment. To examine the role of nonlinearly enhanced heating, lesion stripes were produced in gel by moving the ultrasound transducer. Two different regimes were employed. Variation of duty cycle was combined either with the same transducer velocity and different peak pressure (pulse power or intensity) or with different transducer velocity and the same pulse intensity. Both de-

gassed and nondegassed gel samples were treated to investigate the effect of gas concentration in the gel on lesion development.

In the first set of experiments, a series of seven lesion stripes separated spatially by 5 mm intervals were produced in one gel sample. The transducer was translated upwards at a velocity of 0.5 mm/s. Equal time-average acoustic power of 15 W was retained and duty cycle (dc) gradually increased with corresponding reduction of the peak power from 6.25% dc with 240 W pulse power, to 8.35%, 12.5%, 25%, 50%, 67%, and 100% with 15 W power. For the transducer velocity of 0.5 mm/s, the lowest power level for the cw regime was empirically chosen such that it was just sufficient to induce a symmetric lesion of minimal visible size and without the formation of boiling bubbles. Experiments with static exposures showed that a lesion became visible in 2 seconds and had a characteristic thickness of about 1 mm in 4 seconds at 15 W acoustic power without boiling, similar to the lesions shown in Fig. 3(a). A small increase in power from 15 to 17 W yielded boiling in 2.5 seconds, and doubling of the pulse power with the same average HIFU power of 16.5 W resulted in essentially immediate boiling in single lesion experiments. It was expected, therefore, that all the exposures producing stripes, except the lowest power and the cw regime, would produce observable nonlinear enhancement of heating and boiling. A fundamental difference between static and scanned exposures at a high pulse power level is that, while boiling begins in a very short time in both cases, it cannot persist and distort the lesion for scanned exposures because the focus of the transducer constantly moves away from the boiling site.

A photograph of a gel sample with seven lesion stripes is shown in Fig. 6. HIFU was applied from the front side of the sample. The photo on the left-hand side shows a side view of the largest lesion stripe (numbered 1) that corresponds to the highest pulse power level and lowest duty cycle. The geometry of the cross section of the lesions in the plane perpendicular to the direction of transducer movement is viewed from above of the sample. It is seen that much larger lesions both in length and width were produced for higher peak pressure (pulse power) levels for the same HIFU energy delivered to the sample. The biggest lesion (Fig. 6, number 1) was of 5 mm width and 10 mm length, whereas the smallest lesion (Fig. 6, number 7, only visible from the top) was of 1 mm width and 1.5 mm length. The waveform shape for given ultrasound energy thus dramatically changes resulting lesion size. Lesions produced by mechanically sweeping the transducer over the gel were larger for shorter, stronger pulses, indicating that higher amplitudes and shorter duty cycles cause enhanced heating.

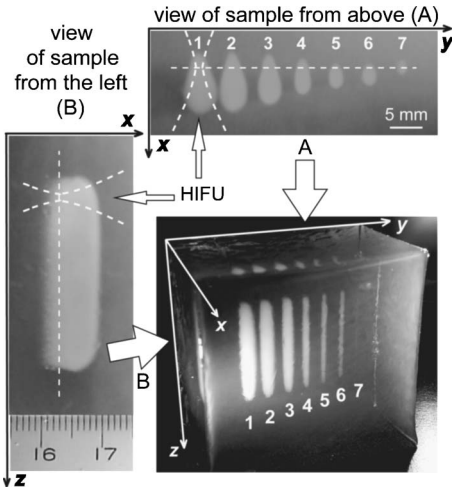


FIG. 6. Lesion stripes formed in degassed gel sample by moving the HIFU transducer upwards with constant velocity (0.5 mm/s), constant time-average acoustic power of the transducer (15 W), and different duty cycle (and peak power) from 6.25% (240 W peak power, 1) to 8.35% (2), 12.5% (3), 25% (4), 50% (5), 67% (6), and 100% (15 W peak power, 7). Ultrasound was applied from the front of the sample as indicated by HIFU arrows. The location of HIFU focal peak is shown by the straight dashed lines, and the general shape of the HIFU beam is shown by the curved dashed lines on the views of the sample from above (A) and from the left (B). Although the average power was held constant for each stripe, the stripes have very different sizes. The highest amplitude, shortest duty cycle pulses created the largest lesion stripe (1).

The geometry of the lesion cross section transformed from the smallest (and symmetrical) lesion located at the focus of the transducer for the lowest peak pressures (7) into the drop-shaped lesions with centroids displaced from the focus toward the transducer. A dashed line on the top photograph corresponds to the peak focal intensity of HIFU and crosses the middle of the smallest lesion [Fig. 6 (lesion 7)]. It is seen that the lesion stripes 1–6, in which we expected boiling, are formed mainly proximally to the focus with a small increase of the lesion size distal to the focus [Fig. 6 (1–3)]. These results are consistent with the observations of lesions in tissue induced at high intensity levels[18,19,33] and support the hypothesis that boiling and corresponding formation of large bubbles effectively reflect HIFU, and is the mechanism for the change in lesion shape from a "cigar" to a "drop" or "tadpole." The stripe produced at the lowest peak pressure (7) was visually observed during the treatment as a moving hazy focal spot which immediately faded in opacity, though it could still be visualized many hours after the exposure.

To investigate the influence of gas content in gel (due to differences in preparation) and different regimes of lesion production, additional experiments were performed using 120 W pulse acoustic power applied to the degassed and nondegassed gels and varying the velocity of the transducer from 0.5 mm/s [Fig. 6 (3)] to 1, 2, and 4 mm/s. As before, the same acoustic energy radiated into the samples was maintained by increasing the duty cycle from 12.5% to 25%, 50%, and 100%, respectively. Under all these exposure conditions, boiling was expected to occur essentially immediately (within a small fraction of a second) given the prior empirical observations. Shown in Fig. 7 are the lesions obtained with 0.5 mm/s (the slowest) and 4 mm/s (the highest) scan velocity in degassed and nondegassed samples. The slower transducer movement resulted in formation of more uniform lesions [Figs. 7(a) and 7(b)]. More visible bubbles formed and remained in nondegassed gels [Figs. 7(b) and 7(d)]. The same differences were observed with the other transducer velocities of 1 and 2 mm/s, but they were more pronounced for the lesions obtained with the slowest and the fastest exposure (Fig. 7). The size of the lesions (size of the

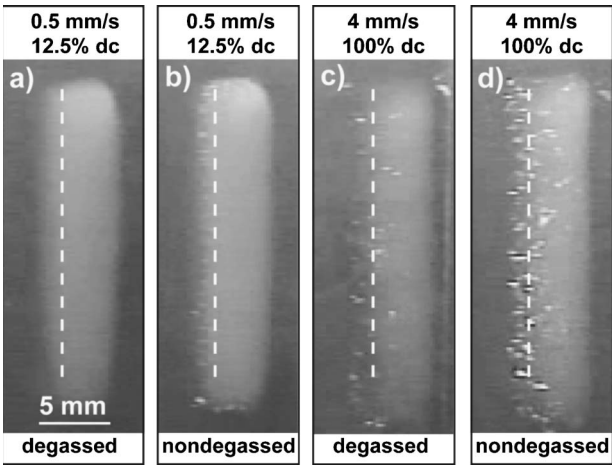


FIG. 7. Comparison of lesion stripes formed in degassed and nondegassed gel samples by moving the HIFU transducer upwards with different scan velocities of 0.5 and 4 mm/s. Insonation was from the right. Peak acoustic power was 120 W, the increase of the transducer velocity was compensated by increased duty cycle from 12.5% to 100% to maintain equal acoustic energy radiated into the sample to create one stripe. Dashed lines indicate the location of the HIFU focal peak. There are more visible bubbles sustained in the nondegassed gels and for faster scan velocity.

diffusive fogged zone) were very similar, however, the distal edge of the lesion was not uniformly treated for nondegassed samples and for higher transducer velocity. The bubbles visible at the distal end of the lesion were located close to the focus, and the proximal end had uniform structure. This suggests the following scenario for lesion development at the high intensity level: boiling starts almost immediately at the focus of the transducer and bubbles reflect ultrasound so that the region of maximum heating shifts toward the transducer. The gel continued to boil at the maximum of the heated region (on axis) but as the lesion grew thicker, boiling bubbles were no longer visible inside the lesion under the side-lighting conditions used in this experiment. In later experiments, a diffuse backlight that made boiling bubbles inside the lesion visible was used with the overpressure apparatus.

C. Overpressure

The results of simulations that accounted for acoustic nonlinearity showed locally enhanced heating leading to temperature rise sufficient for boiling, which correlated well with the experimental observations of single lesions in gel. Conversely, temperature rise obtained with the linear propagation model did not reach boiling at 100 °C. However, both linear and nonlinear simulations were performed with an absorption coefficient measured in gel at a low intensity level[27] and effective absorption at the focus at HIFU intensity might be higher due to the presence of cavitation (HIFU induced microbubbles too small to see), which would result in additional heating in the focal region compounding the effect of nonlinear propagation.

To isolate the effect of acoustic nonlinearity in gel heating from these two bubble-mediated mechanisms, a second set of experiments was conducted with a 2 MHz transducer in a hyperbaric chamber so that greatly elevated static pressure could be applied during HIFU exposure to suppress possible cavitation and increase boiling temperature. Lowering the frequency from 3.5 to 2 MHz was done to increase the role of cavitation by lowering the cavitation threshold at ambient static pressure and to induce slightly larger single lesions, given the larger focal beamwidth obtainable. Acoustic power was chosen so that all the experiments were performed in the presence of cavitation, which was confirmed by preliminary calibration measurements of cavitation threshold in the gel using a passive cavitation detector (PCD). In addition, as the linear absorption coefficient in the gel is nearly proportional to frequency, the difference in heating induced by shock waves versus linear monochromatic waves is stronger for lower fundamental frequencies.[21] Further facilitating observations, the heating process at lower frequency in the linear regime (i.e., in the absence of the effects of nonlinear propagation or cavitation) should be slower, providing more time for the measurements. The changes in lesion development due to nonlinear effects were therefore expected to be more pronounced and more easily observed.

The goal of the initial experiments performed in the overpressure chamber under atmospheric pressure was to visualize characteristic stages of lesion development. Figure 8 represents a sequence of selected video frames for two gel

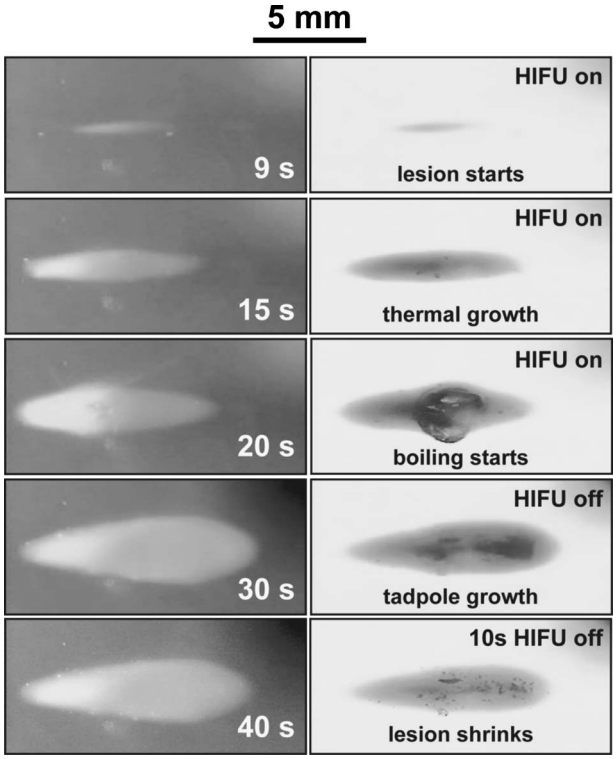


FIG. 8. Sequence of selected video frames illustrating different stages of lesion development in nondegassed gel: lesion inception (9 s after HIFU was on), thermal symmetric growth (15 s), asymmetric tadpole change of lesion shape due to boiling (20 and 30 s), and shrinkage of the lesion soon after HIFU exposure (40 s). Two gel samples were sonicated from the right at ambient static pressure for 30 s with equal acoustic power of 42 W and different illumination either from the left-hand side toward the transducer (frames on the left) or with scattered backlight (frames on the right). Small bubbles of the order of 300 μm were present and visible as dark shadows in the lesion under backlighting conditions at 15 s, but large boiling bubbles of the order of 3 mm were obvious after 20 s.

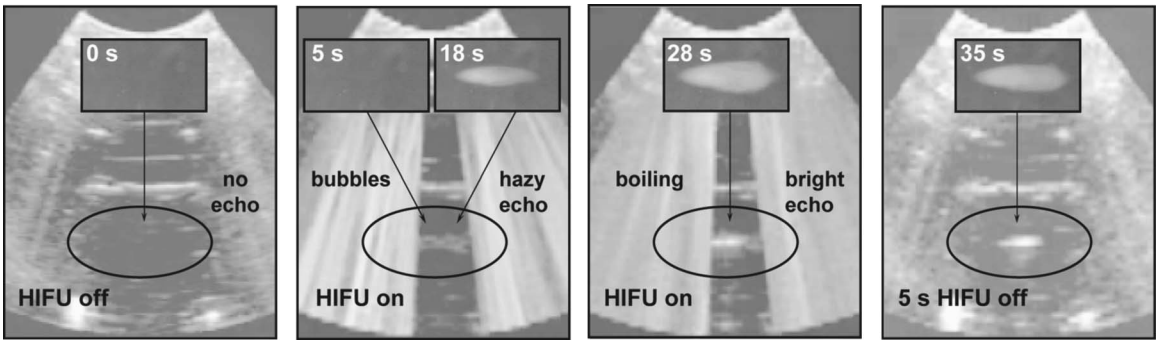


FIG. 9. Simultaneous visualization of the effect of HIFU on nondegassed gel with B-mode imaging and CCD video capture (inset) during 30 s exposure, no overpressure, 36.5 W average acoustic power, and 72% duty cycle. The HIFU transducer was on the right-hand side, lighting was from the left-hand side, and the diagnostic probe was on the back side of the chamber (on the top in the images). The HIFU focal region in the gel is initially hypoechogenic. During HIFU, interference covers all but the center of the image. At the focus, first a weakly brightened zone is seen, and then a strongly echogenic region appears as bubbles become visible in the lesion and the lesion begins to distort dramatically.

samples treated for 30 s with equal acoustic power of 42 W and different illumination, either through the left acrylic window of the chamber towards the HIFU transducer (frames on the left), or with diffuse backlight on the distal window (frames on the right). The power level of the transducer was chosen experimentally to provide clear observations of all stages of lesion development, i.e., lesion inception, symmetric growth, and start of boiling, within the chosen exposure time of 30 s. Side light provided better visualization of the lesion shape and size, and the internal structure of the lesion was better seen when backlit. An LED lamp can be seen (as a blurred zone since the LED was not inside the chamber and outside the camera's depth of field) in the top right corner of each frame indicating when HIFU exposure was in progress. Exposure times are given on the left-hand set of frames and characteristic stages of lesion development in gel are commented on the right-hand set of the frames. The lesion became visible after 9 s of HIFU, it grew symmetrically (15 s) as a thin cigar shape until a large boiling bubble of about 2–3 mm size burst from the middle of the lesion (20 s). The bubble was clearly seen when backlit and was detected as a bright flash when lit from the side (the sidelit and backlit images are taken from different exposure sequences). Boiling bubbles continued to form and break inside the lesion, which started to distort and grow toward the transducer. The heat diffused outward creating the appearance of an evenly denatured lesion in an image obtained with the side light. A typical drop-shaped lesion was formed by the end of the exposure (30 s) with an inhomogeneous bubbly internal structure in the middle (only visible with backlight) and a smoother structure closer to the edges. After HIFU was turned off, the cavities that had formed due to boiling inside the lesion contracted, and the lesion shrank (40 s). The results of this experiment clearly showed the process of boiling in the gel and production of an inhomogeneous internal structure of lesion, very similar to (vaporized) cavities typically observed in overheated tissues (Fig. 4). These results give additional evidence that boiling is responsible for distortion of the lesion shape, and results in rapid growth from the focus towards the transducer. It is important to note that *the experiment was performed in the presence of cavitation* but *no distortion of the lesion was observed before boiling started*. Cavitation was not experimentally detected in this set of experiments because microbubbles cannot be visualized directly, but the acoustic power applied was about three times higher than the cavitation threshold power obtained using PCD measurements in an equivalent setup outside the hyperbaric chamber.

In order to detect microbubble cavitation using a clinically convenient method and compare its appearance with boiling signatures, in the next experiments the results of direct visualization of lesion formation were correlated with diagnostic ultrasound imaging. Microbubbles are strong scatterers of ultrasound, and their presence during HIFU exposure can potentially be seen on B-mode diagnostic ultrasound images, though the absence of B-mode detection cannot guarantee the absence of cavitation.[34] Shown in Fig. 9 is a sequence of B-mode images obtained concurrently with video frames during a 30 s exposure at atmospheric pressure, with 36.5 W average acoustic power, and 72% duty cycle. The gel was heated by HIFU from the right-hand side, lighting was from the left-hand side, and the diagnostic probe was mounted on the back side of the chamber (Fig. 2) oriented with the image plane parallel to the HIFU beam axis so as to view the focal region longitudinally. A weakly echogenic region appeared on the B-mode image immediately after HIFU was turned on and remained in the image through the stages of "no lesion" (5 s) and "cigar-shape thermal lesion" (18 s) observed in the gel. A bright echogenic region was coincident with the start of boiling (start of drop-shaped lesion distortion, 28 s) and persisted after HIFU was off and the lesion shrank (35 s).

These results indicate that cavitation microbubbles may be seen on the B-mode image as a slight increase in echogenicity, even prior to lesion inception. However, this weak echogenicity (enhancement) was not seen in all gel samples exposed under similar conditions. The weak brightening was only perceptible for gels with the most nearly homogeneous and initially scatterer-free structure in the focal area (frame at 0 s), and the possibility remains that in some cases cavitation bubbles were not sufficiently numerous to be detected by a typical observer using a medical B-mode ultrasound scanner. The inhomogeneous structure of tissue would make direct B-mode visualization of the focal region of HIFU preceding boiling even more challenging;[34] nevertheless, more

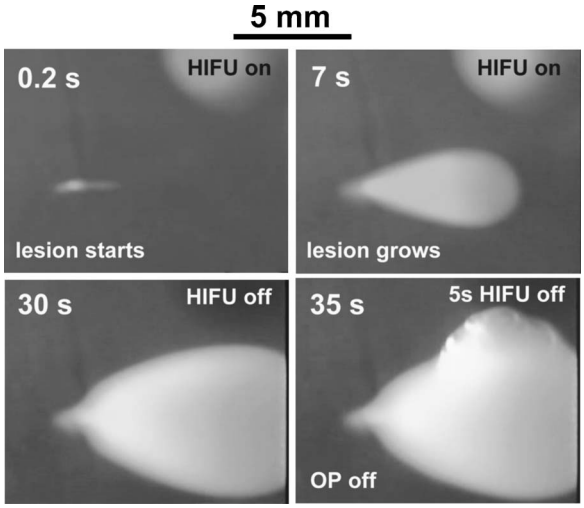


FIG. 10. Development of superheated lesion in degassed gel during 30 s exposure to HIFU under slightly elevated static pressure of 20 bars, and boiling explosion of bubbles from the lesion caused by overpressure (OP) release 5 seconds after HIFU was turned off. The HIFU transducer was on the right-hand side. Acoustic power of the source was 77 W with 100% duty cycle. An LED in the upper right-hand corner of the frames indicates when HIFU was being applied. The explosion of bubbles is seen as the superheated lesion suddenly boils out.

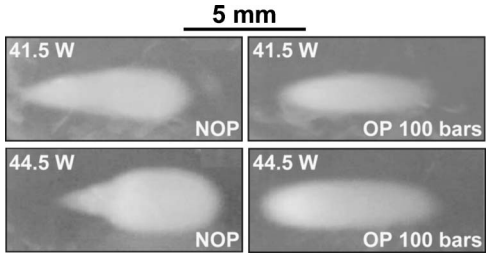


FIG. 11. Comparison of final lesions obtained in degassed gel after 30 s exposure to HIFU (cw regime) under atmospheric static pressure (frames on the left) and under overpressure of 100 bars (frames on the right) with 41.5 and 44.5 W acoustic power chosen below and above the shock formation transition level, respectively. The HIFU transducer was on the right-hand side. Under overpressure the lesions grew symmetrically around the focus to the similar size but different shape than the ones formed without overpressure which were distorted by the effects of boiling bubbles.

sophisticated methods of processing backscattered ultrasound have detected cavitation at the very early stage of HIFU treatment, before boiling.[8,9]

Following boiling, the strongly hyperechogenic region persisted in B-mode observations for more than 1 minute after the treatment, which was consistent with previous numerical and experimental results for tissue.[19,25] The mechanism for dissipation of the echogenicity has been hypothesized to be cooling and shrinkage of the vapor cavities created by boiling, filling of remaining matrix voids with liquid, and dissolution of residual gas. The gel is stiffer and less viscoelastic than tissue and may preserve the shape of cavities formed by bubble growth longer than tissue.

Further experiments were conducted under elevated static pressure. To test the hypothesis that large visible bubbles and ensuing lesion distortion in the gel were due to boiling, and not to HIFU-induced cavitation, one gel sample was treated for 30 s under slightly elevated static pressure of 20 bars and a very high acoustic power of 77 W with 100% duty cycle, exposure conditions which were numerically predicted to lead to boiling temperatures. Because boiling temperature increases with static pressure, a pressurized liquid brought to the boiling point would be superheated with respect to normal atmospheric conditions. Overpressure release after sonication was utilized to confirm boiling inside the lesion. Figure 10 illustrates fast inception of the lesion in gel (0.2 s), lesion growth from the focal spot toward the transducer in a typical tadpole shape caused by boiling (7 s), final lesion shape when HIFU was turned off (30 s), and an explosion of bubbles from the lesion volume when the chamber was depressurized 5 seconds after the end of HIFU exposure (35 s). An LED light in the upper right corner of the frames indicates when the HIFU power was being delivered. Eruptive boiling immediately following decrease in pressure confirmed that the lesioned zone had indeed been raised to the boiling point (at 20 bars over atmospheric pressure).

To examine the effect of overpressure on lesion development and on the onset and characteristics of strong B-mode brightening near the HIFU focal region, gradually increased overpressure levels of 5, 10, 20, and 100 bars were applied in the next series of experiments. Acoustic power from 41.5 to 47 W and 100% duty cycle was used for direct visualization; these levels were chosen because field models indicate that the pressure waveform in the focal region exhibits a rapid transition in this power interval from somewhat distorted (due to nonlinearity) to strongly shocked. In a second sequence of static pressure experiments, 32 W average acoustic power with 50% duty cycle (64 W power during on-time) was used with interleaved B-mode imaging. No weak echoes related to microbubbles were seen on the B-mode images preceding lesion inception and during symmetric lesion growth in any of these overpressure experiments, likely indicating that cavitation was reduced to undetectable levels. Visual detection of boiling bubbles coincided with the appearance of very bright echoes in the focal region in B-mode images, further supporting the hypothesis that only boiling produces such brightening. Higher levels of HIFU, or longer exposures at a given power level, were required to reach boiling for each successive increase in static pressure, consistent with the fact that the boiling point increases with ambient pressure.[35] Moderate overpressure of 5 and 20 bars did not prevent drop-shaped distortion of the lesion shape due to boiling, just as was observed at atmospheric pressure, but it was noted that no lesion distortion occurred with the relatively high overpressure of 100 bars, though boiling was clearly evident. This result is illustrated in Fig. 11, and discussed below.

Final lesions obtained after 30 s of exposure with and without overpressure of 100 bars are presented for comparison in Fig. 11. Lesion inception times (time of first visible discoloration of the gel) were 6 s without overpressure and 14 s with overpressure, at 41.5 W acoustic power; 4 s without overpressure and 10 s with overpressure for 44.5 W. As expected, lesions appeared earlier using higher HIFU power (for a given ambient pressure), and earlier for lower ambient pressure (at a given HIFU power) due to microbubble cavitation which is completely suppressed with 100 bars over-

pressure. Without overpressure, boiling began after lesion inception (at 18 s for 41.5 W, and at 7 s for 44.5 W) and the lesions grew asymmetrically toward the transducer. The lesions appeared substantially later under overpressure and grew symmetrically around the focus without distortion. Perhaps surprisingly, boiling was observed in the 44.5 W/100 bar experiment, but the bubbles were not persistent (they collapsed rapidly) nor was there any tadpole distortion of the lesion shape, as observed in the other cases of lower overpressure in which boiling occurred. The likely explanation is that because the pressure was at 100 bars, the boiling point (reached at the focus) was near 300 °C,[35] and consequently the temperature gradient between the beam axis and the edge of the lesion (near 60 °C) was very high. Bubbles expanding from the axis would cool rapidly, and collapse back into a condensed phase before they persisted long enough to distort the local heating pattern significantly. Similar bubble behavior can be observed on the hot walls of a pan of water that is being heated on a stove, before the average water temperature approaches 100 °C, the walls are much hotter than the boiling point, and effectively superheat water at the boundary and expand explosively only to collapse upon rapid cooling. It is also interesting to note, that in the boiling regime (with 44.5 W) the final lesions obtained after the 30 s exposure with and without overpressure were very different in shape, but the lesion volumes were almost the same (Fig. 11, bottom row). With no overpressure the boiling bubbles make the lesion grow very rapidly at first, scattering the acoustic field and distorting and growing the lesion toward the transducer, and then slowing overall lesion growth rate due to shielding of the focal region where maximum heat deposition would occur.

Final experiments with overpressure were conducted with the goal of isolating the effect of acoustic nonlinearity from the effect of microbubbles in enhancing heat deposition in the gel. The level of overpressure was set to exceed the maximum negative pressure of HIFU which was obtained from the acoustic waveforms simulated numerically for given experimental conditions (Fig. 12). Two different HIFU protocols were applied, one with lower amplitude continuous waves (32 W acoustic power and 100% duty cycle, here called "low amplitude" heating, referring in relative terms to the degree of nonlinearity in the acoustic propagation regime), and another using higher amplitude pulses (64 W pulse acoustic power and 50% duty cycle; "high amplitude" heating regime). A time-average acoustic power of 32 W was maintained for both protocols, expected to produce the same heating results for all cases if nonlinear phenomena were not involved. Images of final lesions formed in gel after 30 s exposures to HIFU under atmospheric pressure and 100 bars overpressure, as well as the focal waveforms modeled for the corresponding two power levels, are presented in Fig. 12. Several observations from this experiment provide insight to the effects of cavitation, acoustic nonlinearity, and the effect of heating to the point of boiling.

In the case of "low amplitude" heating and no overpressure (NOP) the lesion started forming in 15 seconds and grew symmetrically without distortion up to the end of exposure (left-hand frame in the top row of Fig. 12). No visible bubbles were detected. As expected from the use of higher peak pressures, for "high amplitude" heating and no overpressure (NOP) the lesion started much earlier (in 3 seconds), and also grew symmetrically until bubbles appeared in the video image (8 s of exposure) at which time the lesion began to distort in shape and migrate toward the transducer (right-hand frame in the top row).

With overpressure (OP) of 100 bars and "low amplitude" heating the lesion started later than without overpres-

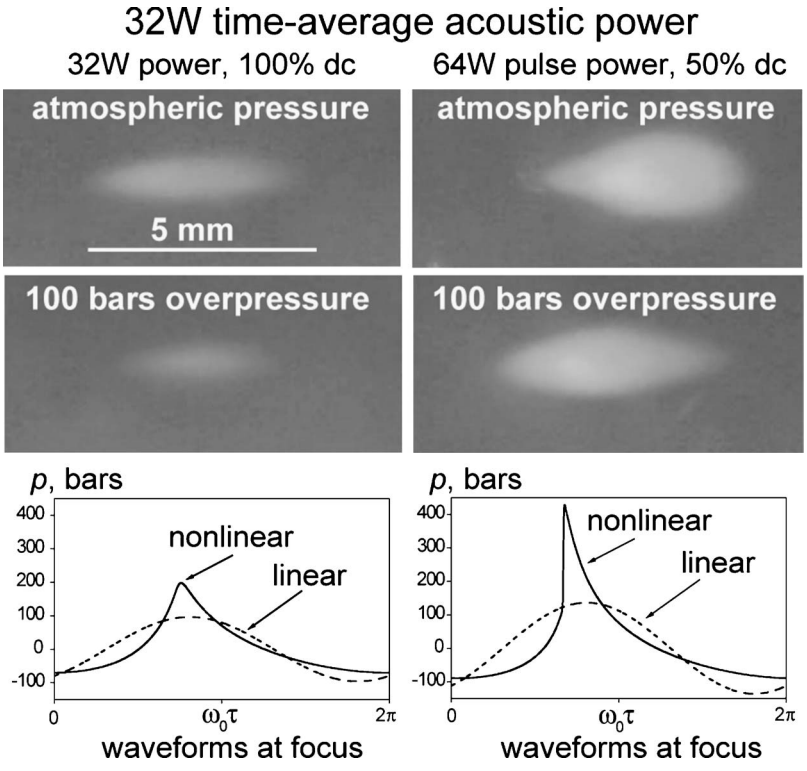


FIG. 12. Final lesions formed in degassed gel under standard atmospheric static pressure (1 bar) and under overpressure (100 bars) after 30 s exposure to HIFU of lower (with 100% duty cycle) and higher (with 50% duty cycle) peak pressure and equal average acoustic power of 32 W. The HIFU transducer was on the right-hand side. Corresponding waveforms are modeled at focus with and without accounting for the effects of nonlinear propagation. The largest change in lesion size is observed by comparing the cases with a change in acoustic pressure, rather than the cases for which only the static pressure was changed. The first comparison isolates the role of acoustic nonlinearity (for high static pressure), whereas the latter comparison isolates the role of cavitation; these observations indicate that nonlinear propagation played a larger role in increasing HIFU heating than did cavitation.

sure (24 seconds of exposure) and had smaller final size (left-hand frame in the bottom row), indicating that, even at reduced acoustic pressure amplitudes, cavitation was clearly contributing to heating (at ambient static pressure). "High amplitude" heating under overpressure resulted in lesion inception in 8 seconds, which was substantially later than without overpressure (again indicating the role of cavitation), but occurred much earlier than in the "low amplitude" OP regime (indicating the importance of acoustic nonlinearity). Again, as shown in Fig. 11 (bottom row), the final size of the "high" OP lesion (right-hand frame in the bottom row) was almost the same as for the "high" NOP lesion, but the shape was different. Although it started later due to the complete suppression of cavitation, the OP lesion grew more rapidly and retained a symmetric shape even after boiling due to strong temperature gradient near the acoustic axis that rapidly cooled expanding vapor bubbles and due to the suppression of boiling bubble growth by overpressure.

The fact that lesion inception times with overpressure were longer than without overpressure for both "low" and "high" heating regimes indicates that cavitation microbubbles not apparent to the eye, but detected in B-mode images (Fig. 9), and measured by PCD, contributed to the enhancement of gel heating. This evidence is consistent with the previously reported results that once the nuclei gas bubbles are activated by HIFU and cavitate, enhanced heating is observed.[14,16,22] Suppression of cavitation by static overpressure has been used previously.[36,37] For the conditions in gel experiments described here, the lesion inception time and final lesion size changed most between the "low" (weakly nonlinear) and the "high" (strongly nonlinear) acoustic intensity regimes, for both ambient and overpressure conditions, one of which completely suppressed cavitation activity. Therefore, we deduce that the effect of acoustic nonlinearity was a major contributor to enhanced heating and accelerated lesion formation, though cavitation also played a role at standard atmospheric pressure.

Many points can be summarized from the overpressure experiments. Overpressure raises the boiling temperature and significantly delayed the appearance of large bubbles—observed simultaneously visually and as a strong rise in echogenicity on B-mode ultrasound—which is consistent with the evidence of the preceding sections that these bubbles (and bright-up) are due to boiling. Explosive large bubble formation upon the release of overpressure was interpretted as boiling resulting from a superheated condition following pressure decrease, and was further evidence that the large, highly echogenic bubbles are the result of boiling, not cavitation. Lesions began to distort visibly only after boiling was evident, and since bubbles continued to appear in the lesion as it migrated and distorted, this result supports previous evidence that reflection from these large bubbles causes the lesion distortion. It is interesting to note that until boiling started, no distortion occurred; yet, cavitation was clearly present in the absence of overpressure, because lesion inception times lengthened substantially with overpressure, for constant HIFU parameters. Thus we conclude that microbubble cavitation, though present, did not result in observable lesion distortion in any of the experiments described here. A subtle, weakly enhanced echogenic region appeared on B-mode images in and near the HIFU focal region at atmospheric pressure prior to boiling, but did not appear when the experiment was repeated with overpressure. From this result we infer that the subtle enhancement in B-mode was evidence of microbubbles; the weak enhancement often appears to blur the original speckle pattern, and gives the impression of a foggy or hazy overlay on the image rather than an increase in brightness. We (and others) have previously seen evidence of microbubbles before detection of a pronounced hyperechogenic region on B-mode. Cavitating microbubbles have yielded accelerated heating in HIFU experiments with these gels and other gel phantoms.[16] In support of enhanced heating due to cavitation we observed that lesions created under overpressure took longer to form than at 1 bar. This result we interpret as evidence that suppression of microbubbles by overpressure reduced heating by cavitation and/or removed the excess acoustic absorption caused by the presence of the bubbles in the gel. In general, the overpressure experiments indicated that, for the range of HIFU intensities that encompassed and exceeded the transition from weakly nonlinear to strongly shocked waveforms, the effect of acoustic nonlinearity was larger than that of microbubble cavitation on the acceleration of lesion formation. Specifically, strongly nonlinear propagation with overpressure (completely suppressing cavitation) ultimately produced the largest lesion, and the lesion remained symmetric and centered at the focus because the shielding effect of boiling bubbles was also suppressed. The HIFU intensities used in ablative therapy are high and acoustic fields are likely to be in the range of strongly nonlinear propagation.

V. CONCLUSIONS

The effects of acoustic nonlinearity, cavitation, and boiling on the dynamics of lesion formation in protein-containing gel phantom exposed to HIFU were investigated experimentally and theoretically in different regimes. Both cavitation and acoustic nonlinearity have been previously proven to enhance heating. One of the goals of this research was to examine the respective roles of these phenomena separately, and especially to use static overpressure to evaluate the role of nonlinear propagation alone in heating of the gel phantom, and to separate the effects of boiling from those of microbubble cavitation.

Our experimental observations confirmed expectations developed using nonlinear acoustic simulations, in particular, the use of high-amplitude tone bursts instead of cw waves of equal acoustic energy leads to accelerated lesion formation and results in larger lesions. Nevertheless, at an ambient pressure that is less than the peak negative pressure of HIFU, cavitation microbubbles also contribute to enhanced heating in gel. This was confirmed here by comparing rates of lesion formation with and without static overpressure that exceeded negative peak pressure and hence completely suppressed cavitation; the cavitation threshold for the gel has been exceeded at standard atmospheric pressure even for weakly nonlinear waveforms. For acoustic power levels near the relatively sharp transition for development of acoustic

shocks near the focus, acoustic nonlinearity was found to be the dominant mechanism for accelerating heating, that is, until boiling occurs. The evidence for this conclusion was that the metrics used for monitoring the acceleration of lesion formation, lesion inception times (shorter for high amplitude pulses), and lesion volumes (larger for high amplitude pulses), were essentially independent of gel gas content or static pressure. The effect of nonlinear acoustic enhancement of lesion production is most pronounced in a narrow zone around the focal axis where shocks form, and thus rapid transverse displacement of the ultrasound transducer (scanning to form lesion stripes) was an excellent approach to visualizing the effect of this enhancement. While static lesion formation is slowed as soon as boiling bubbles appear, scanning permits moving the heat source away from the boiling region before the vapor bubbles have much time to grow and disrupt the highly focused field.

It was shown that several previously observed phenomena were due to boiling in the gel, and not to cavitation, these include greatly increased echogenicity ("bright-up") observed with B-mode ultrasound that is often used to monitor HIFU treatment, distortion of lesions into drop or "tadpole" shape, and rapid lesion growth toward the transducer with associated shielding of the focal zone. Evidence for boiling included the results of nonlinear models of acoustic propagation and heating for the gel experiments, which produced temperatures greater than the boiling point only for those experimental conditions when large bubbles were observed. Furthermore, the sudden creation of persistent and large bubbles (2–3 mm), the similar increase in bubble appearance time with overpressure (that raises the boiling point), and the explosive creation of such bubbles with depressurization of HIFU-heated gel, provide additional support that boiling was responsible. Lesion formation for the gel was essentially instantaneous above 60 °C and always preceded the appearance of visible bubbles.

Highly efficient heating leading to boiling bubble formation was observed above a sharply determined intensity threshold that coincided with the numerically estimated threshold for shock formation in the focal region. This numerical estimate was made without including any factors arising from potential cavitation activity, such as increased values of absorption, for example. Visually observable bubbles first formed in a narrow zone along the HIFU axis (visualized by the lesion) which is consistent with superfocusing by nonlinear wave propagation.[13] The formation of large bubbles which are strong reflectors of ultrasound also coincided with the appearance of a brightly echogenic region on B-mode ultrasound images, and triggered visible change of the shape of ensuing lesion growth and migration toward the transducer. No lesion distortion was observed when boiling bubbles were prevented from growing to a large size by overpressure or before boiling bubbles were observed, even for conditions when microbubble cavitation was occurring. The presence of large bubbles enhances lesion growth due to strong reflections, but it also screens the focal zone of HIFU. With high amplitude-shocked waves and with overpressure used to raise the boiling point and prevent the shielding effect of boiling bubbles, the lesions grew rapidly but symmetrically around the focus to a large size on the same order as that developed with distortion under standard static pressure.

A subtle increase in echo brightness, associated with blurring or a "hazy" appearance, was observed on B-mode images preceding lesion inception and leading up to boiling. This weakly echogenic region was interpretted as evidence of cavitation microbubbles because it was not observed with overpressure, given the same HIFU exposure conditions. With overpressure, lesion inception time was always delayed with respect to the experiments conducted at standard pressure. These observations were interpretted as evidence for cavitation, and cavitation-induced heating, and that the hazy spot is a signature of active cavitation detection using B-mode ultrasound. Since the haze preceded visible lesion formation, more sensitive backscatter processing algorithms to detect and quantify these microbubbles may be useful for targeting and monitoring HIFU treatment.

The experimental results thus show that both cavitation (microbubble dynamics) and nonlinear ultrasound propagation accelerate lesion inception and growth in the gel phantom. By suppressing cavitation with overpressure, and by trading pulse amplitude against duration to accentuate acoustic nonlinearity, nonlinear propagation was observed to be the larger effect. The importance of acoustic nonlinearity in the gel phantom is due to the relatively small value of the attenuation coefficient. Linear attenuation in biological tissue is about 3 times higher than in gel, thus nonlinear gain in heat deposition due to formation of shocks will be lower. However, the results of modeling in tissues at typical HIFU intensities also predict formation of shocks and associated temperature rise above 100 °C, suggesting that nonlinear propagation effects will also be very important for HIFU exposures in biological media.[13,21] The occurrence of boiling in tissue is supported by many experimental observations of vaporized cavities in tadpole-shaped lesions in real tissues.

Nonlinear propagation effects produced in higher peak acoustic pressure regimes, balanced with lower duty cycle to maintain the same time-average transducer power, would improve the efficiency of HIFU treatment by permitting faster transducer scan rates and consequently shorter treatment times. Additional benefits include smaller energy requirements for portable devices and convenient time windows for ultrasonic imaging of treatment when HIFU is off.

ACKNOWLEDGMENTS

The work at the Center for Industrial and Medical Ultrasound (CIMU) at the Applied Physics Laboratory of the University of Washington has been supported by grants from the National Space Biomedical Research Institute (SMS00203) in consortium agreement with the National Aeronautics and Space Administration, the National Science Foundation Grant No. (BES-0002932), and the National Institutes of Health Grant No. (CA83244). The work in the Department of Acoustics at Moscow State University has been supported by the CRDF, RFBR, ONRIFO, and NIH Fogarty. The authors gratefully acknowledge the assistance

of Michael Canney for arranging and performing PCD measurements of cavitation activity in gels.

[1]*Proceedings of the 3rd International Symposium on Therapeutic Ultrasound* (June 22–25, 2003, Lyon, France), edited by J.-Y. Chapelon and C. Lafon.
[2]N. T. Sanghvi and R. H. Hawes, "High-intensity focused ultrasound," Gastrointest Endosc Clin. N. Am. **4**, 383–395 (1994).
[3]S. Vaezy, X. Shi, R. W. Martin, E. Chi, P. I. Nelson, M. R. Bailey, and L. A. Crum, "Real-time visualization of focused ultrasound therapy," Ultrasound Med. Biol. **27**, 33–42 (2000).
[4]M. R. Bailey, V. A. Khokhlova, O. A. Sapozhnikov, S. G. Kargl, and L. A. Crum, "Physical mechanisms of the therapeutic effect of ultrasound," Acoust. Phys. **49**, 437–464 (2003).
[5]F. Wu, W.-Z. Chen, J. Bai, J.-Z. Zou, Z.-L. Wang, H. Zhu, and Z.-B. Wang, "Pathological changes in human malignant carcinoma treated with high-intensity focused ultrasound," Ultrasound Med. Biol. **27**, 1099–1106 (2001).
[6]F. Wu, Z.-B. Wang, W.-Z. Chen, J.-Z. Zou, J. Bai, H. Zhu, K.-Q. Li, F.-L. Xie, C.-B. Jin, H.-B. Su, and G.-W. Gao, "High intensity focused ultrasound for extracorporeal treatment of solid carcinomas: four-year Chinese clinical experience," in *Proceedings of the 2nd International Symposium on therapeutic ultrasound* (July 29–August 1, 2002, Seattle, USA), edited by M. A. Andrew, L. A. Crum, and S. Vaezy, pp. 34–43.
[7]S. Vaezy, M. Andrew, P. Kaczkowski, and L. Crum, "Image-guided acoustic therapy," Annu. Rev. Biomed. Eng. **3**, 375–390 (2001).
[8]P. Phukpattaranont and E. S. Ebbini, "Post-beamforming second-order Volterra filter for pulse-echo ultrasonic imaging," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **50**, 987–1001 (2003).
[9]A. Anand and P. J. Kaczkowski, "Monitoring formation of high intensity focused ultrasound (HIFU) induced lesions using backscattered ultrasound," ARLO **5**, 88–94 (2004).
[10]R. Souchon, L. Soualmi, M. Bertrand, J.-Y. Chapelon, F. Kallel, and J. Ophir, "Ultrasonic elastography using sector scan imaging and radial compression," Ultrasonics **40**, 867–871 (2001).
[11]K. Hynynen, W. R. Freund, H. E. Cline, A. H. Chung, R. D. Watkins, J. P. Vetro, and F. A. Jolesz, "A clinical, noninvasive, MR imaging-monitored ultrasound surgery method," Radiographics **16**, 185–195 (1996).
[12]K. Hynynen, "Demonstration of enhanced temperature elevation due to nonlinear propagation of focused ultrasound in dog's thigh muscle *in vivo*," Ultrasound Med. Biol. **13**, 85–91 (1987).
[13]E. A. Filonenko, G. ter Haar, I. Rivens, and V. A. Khokhlova, "Prediction of ablation volume for different HIFU exposure regimes," in *Proceedings of the 3rd International Symposium on Therapeutic Ultrasound* (June 22–25, 2003, Lyon, France), edited by J.-Y. Chapelon and C. Lafon, pp. 268–274.
[14]K. Hynynen, "The threshold for thermally significant cavitation in dog's thigh muscle *in vivo*," Ultrasound Med. Biol. **17**, 157–169 (1991).
[15]L. A. Crum and G. M. Hansen, "Growth of air bubbles in tissue by rectified diffusion," Phys. Med. Biol. **27**, 413–417 (1982).
[16]R. G. Holt and R. A. Roy, "Measurements of bubble-enhanced heating from focused, MHz-frequency ultrasound in a tissue-mimicking material," Ultrasound Med. Biol. **27**, 1399–1412 (2001).
[17]P. Meaney, M. D. Cahill, and G. ter Haar, "The intensity dependence of lesion position shift during focused ultrasound surgery," Ultrasound Med. Biol. **26**, 441–450 (2000).
[18]F. Chavrier, J. Y. Chapelon, A. Gelet, and D. Cathignol, "Modeling of high-intensity focused ultrasound-induced lesions in the presence of cavitation bubbles," J. Acoust. Soc. Am. **108**, 432–440 (2000).
[19]M. R. Bailey, L. N. Couret, O. A. Sapozhnikov, V. A. Khokhlova, G. ter Haar, S. Vaezy, X. Shi, R. Martin, and L. A. Crum, "Use of overpressure to assess the role of bubbles in focused ultrasound lesion shape," Ultrasound Med. Biol. **27**, 696–708 (2000).
[20]T. G. Muir and E. L. Carstensen, "Prediction of nonlinear acoustic effects at biomedical frequencies and intensities," Ultrasound Med. Biol. **6**, 345–357 (1980).
[21]E. A. Filonenko and V. A. Khokhlova, "Effect of acoustic nonlinearity on heating of biological tissue induced by high intensity focused ultrasound," Acoust. Phys. **47**, 541–549 (2001).
[22]Y. Kaneko, T. Higaki, T. Maruyama, and Y. Matsumoto, "The effect of microbubbles as a heat transducer," in *Proceedings of the 3rd International Symposium on Therapeutic Ultrasound* (June 22–25, 2003, Lyon, France), edited by J.-Y. Chapelon and C. Lafon, pp. 55–60.
[23]L. A. Crum and W. Law, "The relative roles of thermal and nonthermal effects in the use of high intensity focused ultrasound for the treatment of benign prostatic hyperplasia," *Proceedings of the 15th International Congress on Acoustics* (Trondheim, Norway, 1995), pp. 315–318.
[24]N. A. Watkin, G. R. ter Haar, and I. Rivens, "The intensity dependence of the site of maximal energy deposition in focused ultrasound surgery," Ultrasound Med. Biol. **22**, 483–491 (1996).
[25]F. J. Fry, N. T. Sanghvi, R. S. Foster, R. Bihrle, and C. Hennige, "Ultrasound and microbubbles: their generation, detection, and potential utilization in tissue and organ therapy—experimental," Ultrasound Med. Biol. **21**, 1227–1237 (1995).
[26]F. L. Lizzi, D. J. Coleman, J. Driller, R. H. Silverman, B. Lucas, and A. Rosado, "A therapeutic ultrasound system incorporating real-time ultrasonic scanning," *Ultrasonics Symposium* (Institute of Electrical and Electronic Engineers, New York, 1986), pp. 981–984.
[27]C. Lafon, P. Kaczkowski, S. Vaezy, M. Noble, and O. Sapozhnikov, "Development and characterization of an innovative synthetic tissue-mimicking material for high intensity focused ultrasound (HIFU) exposures," *Proceedings of IEEE Ultrasonics Symposium*, 2001, pp. 1295–1298.
[28]O. A. Sapozhnikov, V. A. Khokhlova, T. V. Sinilo, E. A. Filonenko, and L. A. Crum, "Thermal effects of sawtooth waveform HIFU in tissue phantoms," *Proceedings of the 17th International Congress on Acoustics* (Rome, Italy, Sept. 2–7, 2001).
[29]M. O'Donnell, E. T. Janes, and J. G. Miller, "Kramers-Kronig relationship between ultrasonic attenuation and phase velocity," J. Acoust. Soc. Am. **69**, 696–701 (1981).
[30]V. A. Khokhlova, R. Souchon, J. Tavakkoli, O. A. Sapozhnikov, and D. Cathignol, "Numerical modeling of finite amplitude sound beams: Shock formation in the nearfield of a cw plane piston source," J. Acoust. Soc. Am. **110**, 95–108 (2001).
[31]S. Sapareto and W. Dewey, "Thermal dose determination in cancer therapy," Int. J. Radiat. Oncol., Biol., Phys. **10**, 787–800 (1984).
[32]G. ter Haar, "Ultrasound focal beam surgery," Ultrasound Med. Biol. **21**, 1089–1100 (1995).
[33]N. T. Sanghvi, F. J. Fry, R. Bihrle, R. S. Foster, M. H. Phillips, J. Syrus, A. Zaitsev, and C. Hennige, "Microbubbles during tissue treatment using high intensity focused ultrasound," *Proceedings of IEEE 95 UFFC Symposium*, 1995, pp. 1249–1253.
[34]B. A. Rabkin, V. Zderic, and S. Vaezy, "Hyperecho in ultrasound images of HIFU therapy: involvement of cavitation," Ultrasound Med. Biol. **31**, 947–956 (2005).
[35]*Handbook of Chemistry and Physics*, edited by R. C. Weast (CRC Press, Boca Raton, FL, 1985), pp. D186–D187.
[36]W. J. Fry, D. Tucker, F. J. Fry, and V. J. Wulff, "Physical factors involved in ultrasonically induced changes in living systems: II. Amplitude duration relation and the effect of hydrostatic pressure for nerve tissue," J. Acoust. Soc. Am. **23**, 364–368 (1951).
[37]P. P. Lele, "Effects of ultrasound on ′solid' mammalian tissues and tumors *in vivo*," *Ultrasound: Medical Applications, Biological Effects and Hazard Potential* (Plenum, New York, 1986), pp. 275–306.

Information entropy of humpback whale songs[a)]

Ryuji Suzuki[b)]
Department of Electrical and Computer Engineering, University of Massachusetts Dartmouth, North Dartmouth, Massachusetts 02747-2300

John R. Buck[c)]
Department of Electrical and Computer Engineering and School for Marine Science and Technology, University of Massachusetts Dartmouth, North Dartmouth, Massachusetts 02747-2300

Peter L. Tyack
Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543

(Received 2 December 2005; accepted 6 December 2005)

The structure of humpback whale (*Megaptera novaeangliae*) songs was examined using information theory techniques. The song is an ordered sequence of individual sound elements separated by gaps of silence. Song samples were converted into sequences of discrete symbols by both human and automated classifiers. This paper analyzes the song structure in these symbol sequences using information entropy estimators and autocorrelation estimators. Both parametric and nonparametric entropy estimators are applied to the symbol sequences representing the songs. The results provide quantitative evidence consistent with the hierarchical structure proposed for these songs by Payne and McVay [Science **173**, 587–597 (1971)]. Specifically, this analysis demonstrates that: (1) There is a strong structural constraint, or syntax, in the generation of the songs, and (2) the structural constraints exhibit periodicities with periods of 6–8 and 180–400 units. This implies that no empirical Markov model is capable of representing the songs' structure. The results are robust to the choice of either human or automated song-to-symbol classifiers. In addition, the entropy estimates indicate that the maximum amount of information that could be communicated by the sequence of sounds made is less than 1 bit per second.
[DOI: 10.1121/1.2161827]

PACS number(s): 43.80.Ka [WWA] Pages: 1849–1866

I. INTRODUCTION

Payne and McVay (1971) first analyzed the structure of the songs of humpback whales (*Megaptera novaeangliae*). Their analysis relied on human judgments both to determine whether one sound element was identical to another and to determine whether the song fit the hypothesized hierarchical song structure. Subsequent studies (Guinee and Payne, 1988; Miller *et al.*, 2000; Noad *et al.*, 2000; Payne *et al.*, 1983) also relied on human observers for these evaluations. This paper presents methods to remove these subjective judgments by employing an automated classifier to group the units and applying objective information theory techniques to study the song structure.

The rest of this section reviews the previously proposed structure for humpback whale songs and provides an overview of the elements of information theory necessary for this research. Recognizing that information theory falls outside the purview of most animal bioacousticians, the most important theorems and concepts are interpreted in the context of animal communications studies.

A. Songs of humpback whales

Most humpback whales have an annual migratory pattern, breeding in subtropical latitudes during winter, and migrating to high latitude waters to feed in the summer. The vocalizations produced by humpback whales during these feeding and breeding seasons differ greatly. The feeding calls involve a few simple sound patterns produced in simple sequences (D'Vincent *et al.*, 1985), whereas whales produce complex songs during the breeding season. The term *song* is used in animals, such as songbirds and whales, to describe an acoustic signal that involves a wide variety of sounds repeated in a specific sequence.

Humpback songs consist of a sequence of discrete sound elements, called *units*, that are separated by silence. Each song contains a complicated series of more than 12 different units. These units cover a wide frequency range (30–3000 Hz), and consist of both modulated tones and pulse trains. Payne and McVay (1971) proposed a hierarchical structure for humpback song. A song is a sequence of *themes*, where a theme consists of a *phrase*, or very similar phrases, repeated several times. A phrase is a sequence of several units. The song is repeated many times with considerable accuracy to make a *song session*. The reported range of song duration is from 7 to 30 min (Payne and McVay, 1971) or from 6 to 35 min (Winn and Winn, 1978). Winn and Winn (1978) also reported the maximum duration of

[a)]Parts of this paper were presented at the ASA/EAA/DEGA meeting held in Berlin, Germany in March 1999.
[b)]Current address: Speech and Hearing Bioscience and Technology, Harvard-MIT Division of Health Science and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139.
[c)]Electronic address: johnbuck@ieee.org

observed song sessions to be 22 h. Throughout this paper, we use song length to indicate the number of units in a song, and song duration to indicate the number of minutes the song lasts.

All whales in a population are singing the same or very similar songs at a given time, although whales within hearing range do not coordinate to sing the same part of the song at the same time. The songs within a population gradually change over time, so that after several singing seasons few elements of the song have been preserved (Payne *et al.*, 1983). Several reviewers believe that the speed and pervasiveness of this change indicates that singing whales must learn each sound unit and the sequence order that make up a full song (Janik and Slater, 1997; Tyack and Sayigh, 1997). Guinee and Payne (1988) suggested that this song evolution presents a difficult learning and memory task. They proposed that humpback whales increase the redundancy of parts of phrases between adjacent themes as a mnemonic aid, and they found that this redundancy was more common in songs with larger numbers of themes where more material would have to be remembered.

This paper analyzes humpback song with a suite of techniques to determine whether the song contains mathematical properties consistent with hierarchical structure. In this paper, the term *unit* represents the smallest sound element separated by silence, as defined by Payne and McVay (1971), and the term *song* denotes a sequence of units, but the analysis methods developed do not assume the hierarchical syntax proposed by Payne and McVay (1971). A humpback whale song *classifier* is defined as a method of extracting individual units from a continuous sound recording and converting them into discrete symbols (i.e., A, B, C,…) that represent the particular types of sound, preserving the sequence order. In so doing, the classifier absorbs minor variations between sounds by lumping them together as a single symbol. The classifier assigns a new symbol when it encounters a sound element that is deemed sufficiently different from all previously encountered units. Each symbol has no semantics other than representing the particular type of unit. Winn and Winn (1978) performed classification by assigning "approximate phonetic terms" such as "moan," "snore," "cry," and "chirp" to the units. Our analysis of song structure is based on a sequence of units represented by abstract symbols and, therefore, it ignores these acoustic features, duration, and other details of units and gaps between units.

In this paper, the structural constraints in the symbol sequence are referred to as a *syntax* or *grammar*, with no relation to semantics that may or may not exist. This differs from formal language theory where a syntax denotes a set of rules that generates all legal sentences but no illegal ones based on the language's grammar. This paper focuses on the question of which class of syntactic models are the simplest and most accurate for generating sequences of units that match the sequences in humpback songs. Like Payne (1995), no claims positive or negative are made about the semantic content of the songs.

Our analysis addressed three questions. First, what is the quantitative measure of the information conveyed by the songs? Second, can any of the common stochastic models used in animal communication studies reasonably approximate the statistical features of the songs? Third, does a hierarchical model provide the best match to the structure of the songs?

B. Information theory

This section provides an overview of information theory as it pertains to the present study, including several topics not previously employed to analyze animal communications. The primary focus of the section is methods of entropy estimation, but it is first necessary to provide a framework within which these estimates can be properly interpreted. Information theory studies an information *source* $\mathcal{X}$, which produces a sequence, or stream, of discrete symbols. The sequence produced is not assumed to be purposeful or meaningful. Information theory provides techniques to analyze the characteristics of the source, such as the structure of its output, without knowledge of the semantics of the source or its output.

1. Coding bound

At each time i the source $\mathcal{X}$ produces a symbol x_i, for $i=1,2,3,\ldots$. The notation x_i^j represents the subsequence $(x_i, x_{i+1}, \ldots, x_j)$, $i \leqslant j$. An output symbol is called a *letter*, and the set of all possible symbols an *alphabet*, denoted by set $\mathcal{A}$. The entire output sequence is called a *message*, denoted by x_1^∞. Again, there is no assumption that the letters, message, or alphabet are meaningful.

An encoder converts a message to a binary sequence, or *code*, from which a matching decoder can uniquely and exactly reproduce the original message.[1] An encoder is denoted by $f(\cdot)$. Let $\|f(x_1^\ell)\|$ denote the code length of x_1^ℓ, i.e., the number of binary digits, or *bits*, produced by f to encode the message x_1^ℓ of length ℓ.

Definition 1 The *coding rate* of the encoder f for a given sequence x_1^∞ is defined as

$$R_f(x_1^\infty) = \lim_{\ell\to\infty} \sup \frac{\|f(x_1^\ell)\|}{\ell}, \tag{1}$$

that is, the average number of bits per symbol required to code the sequence in the limit of an infinitely long sequence.

There is a lower bound on the coding rate, and there is always some encoder that attains this bound.

Theorem 1 (Coding theorems) For a given message x_1^∞, there is some R' such that

$$R_f(x_1^\infty) \geqslant R' \tag{2}$$

holds for any encoder f. Moreover, there exists an encoder f_{opt} which satisfies Eq. (2) with equality, that is

$$R_{\text{opt}}(x_1^\infty) = R'. \tag{3}$$

Equation (3) is called the coding theorem and Eq. (2) is called the converse coding theorem.

Asymptotically optimal encoders satisfy Eq. (3), achieving the smallest possible code rate in the limit of an infinitely long data sequence. Conversely, encoding x_1^∞ at a rate R less than R' guarantees that x_1^∞ cannot be faithfully recovered from the code $f(x_1^\infty)$, and information will be lost. Conse-

quently, R' is the per letter information contained in the message, and is closely related to the information entropy, as will be shown shortly.

2. Entropy, definition and properties

A source $\mathcal{X}$ may also be considered as a stochastic process, or *process*, producing a sequence of outputs, or a message, controlled by the source's *probability law*. The source's probability law dictates the probability of seeing any given message segment x_i^j for any $1 \leqslant i \leqslant j$ as the output of the source. Let X_i be a random variable representing the source output at time i, X_i^j denote $(X_i, X_{i+1}, \ldots, X_j)$, $i \leqslant j$, and $\mathcal{A}$ an alphabet whose size is $|\mathcal{A}|$. A particular sample output of the random variable is denoted by the lower case letter corresponding to the random variable, e.g., x_i^j. Let $p(x)$ denote $\Pr\{X=x\}$.

Sources where all joint probability distributions are time invariant are called *stationary*. That is, for every possible subsequence $x_i^{i+\ell-1}$ of length $\ell \geqslant 1$, $p(x_i^{i+\ell-1})$ remains the same for all i. Sources where all realizations of the output possess the same statistical properties are called *ergodic*. If the source $\mathcal{X}$ is a stationary ergodic source, it behaves in the same manner in a statistical sense every time the source is started, and also at any given time during the operation. Thus, for a stationary ergodic source, the statistical properties of one long output can be generalized to the properties of the source itself.

A source has memory if the ith output X_i is not statistically independent of all past outputs. If the current output depends *only* on the previous k outputs, the source is a kth-order Markov source.[2] A Markov source where it is possible to reach any state from any state with positive probability in a finite number of steps is said to be irreducible (p. 61, Cover and Thomas, 1991). The Doeblin condition[3] describes a broader class of sources that is less restrictive than Markov sources. The Doeblin condition limits the largest influence that the past may have on the present, while still allowing the influence of the previous outputs from arbitrarily far in the past. Informally, a source satisfying the Doeblin condition blurs the memory of the past output after k symbols; no matter what the output sequence has been up until the current time, all outputs are possible again k symbols into the future. Any irreducible Markov source of a finite order satisfies the Doeblin condition. Also, the Doeblin condition is satisfied if there is an arbitrarily small positive probability of observation error. Thus, almost any practical experimental observation of a stationary source satisfies this condition.

The classes of common information sources can be organized in a list of increasingly strong restrictions, and each class in this list includes all of the subsequent classes as a subset. This list of classes is: (1) Sources which are not stationary nor ergodic, but are governed by a consistent probability law; (2) stationary ergodic sources; (3) stationary ergodic sources that satisfy the Doeblin condition; (4) stationary ergodic irreducible Markov sources of finite order; and (5) independently identically distributed (i.i.d.) sources. Classes appearing later in this list are more restricted in their applicability than earlier ones due to the strong assumptions imposed on the sources in the later class. Note that within the Markov sources, a lower-order Markov source is more restrictive than a higher-order one.

The following discussion of entropy and related source properties summarizes the cumulative contributions of numerous authors. In the interest of brevity, only the strongest results are presented below. The Appendix outlines the progression of these results. Except where noted otherwise, these properties hold for each source class described in the list above.

Definition 2 The *entropy* of a source is defined as

$$H(X_1^\infty) = \lim_{\ell \to \infty} -\frac{1}{\ell} \sum_{x_1^\ell} p(x_1^\ell) \log p(x_1^\ell), \tag{4}$$

where the summation is over every possible subsequence of length ℓ. Throughout this paper, $\log x$ means $\log_2 x$, and $0 \log 0$ is 0. The units of entropy are bits. The entropy of a source is the minimum average number of bits per symbol necessary to encode its messages.

Theorem 2 (Entropy as the coding bound). In the limit, the coding rate of an asymptotically optimal encoder meets the entropy of the source. That is,

$$R_{\text{opt}}(x_1^\infty) = H(X_1^\infty), \quad \text{with probability one.} \tag{5}$$

This theorem allows another interpretation: The entropy of a source represents the average amount of information per symbol that the source transmits in its messages over a long symbol sequence. Given two messages of length ℓ from alphabet $\mathcal{A}$, x_1^ℓ and y_1^ℓ, if the encoding of x_1^ℓ is shorter than that of y_1^ℓ, i.e., $\|f_{\text{opt}}(x_1^\ell)\| < \|f_{\text{opt}}(y_1^\ell)\|$, then x_1^ℓ is said to be more *compressible* than y_1^ℓ. Redundancy in a message can be measured in terms of entropy, and interpreted in terms of compressibility, in light of Theorem 2. A lower entropy of a source implies that the message from the source is more compressible and more redundant. Consequently, entropy is a measure of the average redundancy of the messages from a source.

From the receiver's point of view, entropy is the average measure of *a priori* uncertainty about each successive letter of the source output, and hence the amount of information received equals the amount of uncertainty removed. Equivalently, since a source with little uncertainty is very predictable, entropy decreases with increased predictability.

MacKay (1972) observed "one has no prior reason to regard H as a more biologically significant measure of the *information received* by that organism than, say, the total number or duration of the signals exchanged," (MacKay, 1972, p. 11). However, a properly obtained estimate or upper bound on H is an upper bound of the information received by an organism. Moreover, recent advances in information theory offer new approaches to biological problems beyond the limited methods that MacKay considered.

Entropy is also a measure of the structural constraints and complexity of a source.

Property 1 Additional constraints in the structure of an information source decrease the entropy of that source.

The size of the alphabet restricts the maximum possible entropy to be $H_{\max}=\log|\mathcal{A}|$, which can be attained if and only if each output symbol is independent and uniformly distributed. For a source with entropy $H_{\max}=\log|\mathcal{A}|$, there are $|\mathcal{A}|^{\ell}=2^{\ell H_{\max}}$ possible sequences of length ℓ which may be observed, each with equal probability $|\mathcal{A}|^{-\ell}=2^{-\ell H_{\max}}$. If the output symbols are not independent and uniform, i.e., the source has structural constraints, the entropy H is less than $H_{\max}$. As noted earlier, decreased source entropy implies increased redundancy in the output streams, and this redundancy is quantified as

$$\rho=(H_{\max}-H)/H_{\max}, \tag{6}$$

where $H_{\max}$ is computed for the same alphabet size as the source (Shannon, 1948). Thus, the redundancy ρ is between 0 and 1. Structural constraints imply that some sequences are much more probable than others. The highly probable subset of sequences is called the *entropy-typical set*, and the probability of the other atypical output sequences appearing is very small. This is formally represented by the following theorem.

Theorem 3 (Entropy theorem).

$$\lim_{\ell\to\infty}-\frac{1}{\ell}\log p(x_1^{\ell})=H \quad \text{with probability 1}. \tag{7}$$

Note that this theorem concerns a long individual sequence, and not the average of all sequences. Rearranging Eq. (7) leads to:

Theorem 4 (Typical set). For a positive integer ℓ and $\epsilon>0$, define the entropy-typical set to be

$$\mathcal{T}(\ell,\epsilon)=\{x_1^{\ell}:2^{-\ell(H+\epsilon)}\leqslant p(x_1^{\ell})\leqslant 2^{-\ell(H-\epsilon)}\}. \tag{8}$$

Then, there is a sequence length $\ell'(\epsilon)$ such that for all $\ell>\ell'(\epsilon)$,

$$\Pr\{\mathcal{T}(\ell,\epsilon)\}>1-\epsilon \tag{9}$$

and

$$(1-\epsilon)2^{(H-\epsilon)\ell}\leqslant|\mathcal{T}(\ell,\epsilon)|\leqslant 2^{(H+\epsilon)\ell}. \tag{10}$$

That is, for a small ϵ and sufficiently large ℓ: (1) Almost all of the observed sequences belong to the typical set [Eq. (9)]; (2) all sequences in the typical set are roughly equally likely to occur with probability close to $2^{-H\ell}$ [Eq. (8)]; and (3) the size of the typical set is approximately $2^{H\ell}$ [Eq. (10)]. The atypical set $\mathcal{T}^c(\ell,\epsilon)=\{x_1^{\ell}:x_1^{\ell}\notin\mathcal{T}(\ell,\epsilon)\}$ contains sequences that are improbable or do not occur. When H is significantly smaller than $H_{\max}$, the set of the typical sequences is only a tiny subset of the set of all possible sequences, $\mathcal{A}^{\ell}$, and the size of the atypical set $|\mathcal{T}^c(\ell,\epsilon)|$ is very large. Summarizing this theorem:

Property 2 Stationary ergodic sources with entropy H typically produce $2^{H\ell}$ approximately equiprobable sequences of length ℓ.

Therefore, for a given message length, a source with a larger entropy produces a greater variety of alternative messages than a source with a smaller entropy. Note that this statement concerns only the entropy H, and that the alphabet size $|\mathcal{A}|$ is not a direct factor in the number of typical messages.

Example For the English alphabet of 26 letters plus a space ($|\mathcal{A}|=27$), consider two processes producing very long sequences. The first source is a monkey typing on a keyboard such that each letter is equally probable, and the second source is an English writer. We regard both of these as information sources, and take subsequences of an arbitrarily chosen length 150 from each source output, and call each of them a sentence. The monkey process has the maximum entropy possible for the alphabet size, $H=H_{\max}=\log 27\approx 4.755$ bits, producing all $27^{150}=2^{150\log 27}\approx 5\times 10^{214}$ possible sentences with equal probability. This is the size of the typical set for the "monkey typing" source. Cover and King (1978) estimated the source entropy of English to be approximately 1.3 bits. Thus, English writers typically produce $2^{1.3\times 150}\approx 5\times 10^{58}$ sentences. The size of this typical set is $2^{(\log 27-1.3)150}\approx 10^{156}$ times smaller than the typical set of the monkey process, which produces all possible sentences. The output of the English writer has structural constraints in the form of lexicographical and grammatical rules and the context of the story, limiting the size of the English typical set to be a tiny portion of the size of the monkey typing source typical set.

These interpretations of entropy provide the basis for studying the structure of humpback whale songs. The direct application of Eq. (4) for entropy estimation is difficult for practical problems with a finite set of observations, because the *true* probability law for the sequence of the units in humpback song is not known. Instead, $p(\cdot)$ must be estimated from the available observations, which are necessarily finite in length. One common method is to use $\hat{p}(\cdot)$, the observed *relative frequencies* of events, or the *empirical distribution*. The true distribution and empirical distribution are conceptually different quantities, and without *a priori* knowledge of the source, there is no guarantee that the empirical distribution $\hat{p}(\cdot)$ from an observed output sequence is within a given tolerance of the true probability $p(\cdot)$.

As in any model-based estimation problem, the estimation of the source entropy generally requires: (1) Establishing and justifying a stochastic model for the source, (2) estimating the model parameters (the probability mass function), and then (3) estimating the entropy of the source. Two popular source models used for Step (1) are the i.i.d. model and the empirical Markov model. These models are often assumed under speculation or without justification in the study of animal communications. For example, Beecher (1989) used i.i.d. entropy estimates as a measure of the information embedded in signature calls, implicitly assuming that each call is statistically independent and identically distributed. Gentner and Hulse (1998) used Markov models of varying order with no justification that the sources analyzed fit these models. Another approach to estimation problems is the use of model-free, or nonparametric, methods. In the nonparametric approach, the selection of a model is unnecessary, and therefore the approach is more universally applicable.

Instead of justifying a specific model, we chose to per-

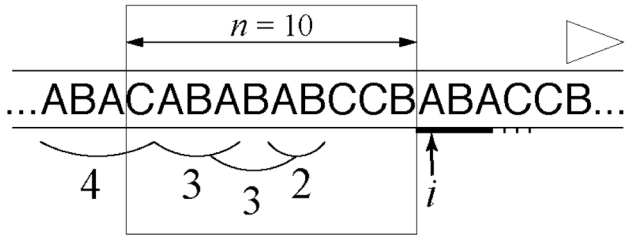


FIG. 1. Example of the sliding window match length. A section of the sample sequence with current index i and window of size $n=10$ are shown. The match length at this index is $L_i(10)=\max\{3,3,2\}=3$. Note that the longer match of ABAC with length 4 is not counted because it begins before the start of the window.

form the entropy estimation with three methods: (1) Estimation with an i.i.d. model; (2) estimation with an empirical Markov model of order 1; and (3) a nonparametric method. Each model's ability to approximate the structure embodied in the humpback songs can be assessed from the resulting entropy estimates using Property 3 (Sec. I B 4).

3. Model-based entropy estimation

The entropy of the i.i.d. and Markov models can be simplified from the general expression in Eq. (4). The model-based estimators result from substituting the observed empirical distributions $\hat{p}(\cdot)$ for the unknown probability mass function (PMF).

Definition 3 (i.i.d. model). The source entropy of an i.i.d. process is estimated from a sample sequence x_1^ℓ of length ℓ by

$$\hat{H}_0 = -\sum_{a\in\hat{\mathcal{A}}} \hat{p}(a)\log \hat{p}(a), \tag{11}$$

where the empirical distribution of the individual letters is

$$\hat{p}(a) = \frac{|\{i{:}x_i = a, i \in [1,\ell]\}|}{\ell}, \tag{12}$$

and the estimated alphabet $\hat{\mathcal{A}}$ is

$$\hat{\mathcal{A}} = \{x_i{:}i \in [1,\ell]\}. \tag{13}$$

That is, the empirical distribution for a letter $\hat{p}(a)$ is the number of appearances of that letter in the sequence divided by the length of the sequence, and the estimated alphabet $\hat{\mathcal{A}}$ is the set of all symbols appearing in the data sequence. For example, for the sequence segment of length $\ell=19$ in Fig. 1, $\hat{p}(\mathrm{A})=7/19$, $\hat{p}(\mathrm{B})=7/19$, and $\hat{p}(\mathrm{C})=5/19$, for the i.i.d. model. The resulting entropy estimate for this unrealistically short example sequence would be $\hat{H}_0 \approx 1.57$ bits. This example is meant for illustrative purposes only, and in practice $\hat{H}_0$ should only be computed for much longer sequences.

The i.i.d. model ignores all statistical dependencies between occurrences of letters by treating them as conditionally independent, and assumes that they are drawn from an identical distribution. The presence of memory in a statistical sense means that the present state or output is at least partially a consequence of the past states. The i.i.d. model fails to represent any effect of memory; hence, it is sometimes called the stationary memoryless model. Such a model is often used as a measure of individual divergence in ecology (Good, 1953; Patil and Taillie, 1982; Peet, 1974). Equation (12) relies on the strong law of large numbers to ensure that the empirical distribution, $\hat{p}(\cdot)$, converges to the true distribution with probability one as $\ell\to\infty$. Therefore, the i.i.d. condition is crucial for this model to be valid, i.e., for $\hat{H}_0$ to converge to the true H.

The simplest source model with memory is the first-order Markov model, where only the immediately previous output influences the current output through the conditional probability law.

Definition 4 (Markov model). The source entropy of a Markov model of order $k=1$ is estimated from a sample sequence by

$$\hat{H}_1 = -\sum_{a_1,a_2\in\hat{\mathcal{A}}} \hat{p}(a_1,a_2)\log \hat{p}(a_2|a_1), \tag{14}$$

$$= -\sum_{a_1,a_2\in\hat{\mathcal{A}}} \hat{p}(a_1,a_2)\log \hat{p}(a_1,a_2) - \hat{H}_0, \tag{15}$$

where the empirical distribution of the conditional and joint probability are obtained by counting the occurrences of each pattern:

$$\hat{p}(a_2|a_1) = \frac{|\{i{:}x_{i-1}=a_1, x_i=a_2, i\in[2,\ell]\}|}{|\{i{:}x_{i-1}=a_1, i\in[2,\ell]\}|}, \tag{16}$$

and

$$\hat{p}(a_1,a_2) = \frac{|\{i{:}x_{i-1}=a_1, x_i=a_2, i\in[2,\ell]\}|}{\ell-1}. \tag{17}$$

Equation (16) is equivalent to counting what fraction of the appearances of a_1 are immediately followed by a_2. For the example in Fig. 1, $\hat{p}(\mathrm{B}|\mathrm{A})=5/7$, $\hat{p}(\mathrm{C}|\mathrm{C})=2/5$, but $\hat{p}(\mathrm{A}|\mathrm{B})=5/6$ since we cannot count the final B because we do not know what symbol follows. Equation (17) is equivalent to counting each possible pair of letters, then dividing by $\ell-1$, the total number of pairs. Referring to Fig. 1, $\hat{p}(\mathrm{A},\mathrm{B})=5/18$, $\hat{p}(\mathrm{B},\mathrm{A})=5/18$, $\hat{p}(\mathrm{A},\mathrm{C})=2/18$, $\hat{p}(\mathrm{C},\mathrm{A})=1/18$, etc…. Substituting appropriate empirical distributions for this unrealistically short segment yields the estimate $\hat{H}_1 \approx 0.98$. Again, this example is only meant for illustration, and in practice $\hat{H}_1$ should only be computed for much longer sequences. More generally, a kth-order Markov model requires estimates of the joint and conditional probabilities from observed blocks of length $k+1$.

Equations (16) and (17) rely on the ergodic theorem to ensure that the empirical joint distribution of the overlapping blocks of length 2 contained in a single length-ℓ block sample sequence converge to the true distribution with probability one as $\ell\to\infty$. It is crucial for the validity of this model that the stationary ergodic and Markov properties be maintained throughout the entire sample sequence. If these assumptions are not valid, the resulting entropy estimates

obtained by using this model may be significantly flawed. Secs. I B 5 and I B 6 discuss additional difficulties in the practical application of Markov model entropy estimates even when all the model assumptions are valid.

4. Model-free entropy estimation

As noted in Sec. I B 1, the encoder f is a data compressor; an encoder compresses the sequence x_1^ℓ to a code of length $\| f(x_1^\ell) \|$ bits, which preserves all of the information in the original sequence and is usually more compact than the original sequence. A *universal* data compressor is an encoder that does not require *a priori* knowledge of the source's probability law (Kolmogorov, 1965). An asymptotically optimal data compressor is one where the average code length approaches entropy as the sequence length increases [Eq. (5)]. Ziv and Lempel (1977) developed the first universal and asymptotically optimal data compression algorithm, called LZ'77 (Wyner and Wyner, 1994; Ziv and Lempel, 1978a). Theoretically, one can use LZ'77 as the encoder $f(\cdot)$ in Eq. (1) with a *long* observed sequence and then consider the resulting code rate $R_f(x_1^\ell)$ as an estimate of the entropy of the source. In practice, this estimator converges to H too slowly in the sequence length ℓ to be useful in most situations. No commonly employed asymptotically optimal universal data compression algorithm converged quickly enough to be useful for the data lengths available in this study.

The nonparametric entropy estimator presented in this section resulted from a performance analysis of the LZ'77 encoder. This estimator provides reliable and robust entropy estimates, and converges to the true entropy H much faster than the LZ'77 approach described in the previous paragraph. Even if the source is nonstationary, or the sample sequence is not long enough to achieve asymptotic convergence, the resulting entropy estimates are an upper bound on the source entropy, and thus is an upper bound on the amount of information transmitted (Theorem 2).

The nonparametric entropy estimate is computed from the *match length* statistics.

Definition 5 (Match length). Match length $L_i(n)$ is defined for a fixed window size n and current sequence index i by

$$L_i(n) = \max\{L : x_i^{i+L-1} = x_{i-k}^{i-k+L-1}, k \in [1,n]\}. \tag{18}$$

Figure 1 illustrates an example of computing the match length. Consider the sequence of symbols shown. The match length is the length of the longest string starting at the present sample x_i that matches a string which begins within the window of n symbols immediately preceding x_i. In Fig. 1, $n=10$, and the longest match $L_i(10)=\max\{3,3,2\}=3$.

Theorem 5 For stationary ergodic sources, the relation between the match length and source entropy is

$$\lim_{n\to\infty} \frac{L_i(n)}{\log n} = \frac{1}{H}, \quad \text{with probability one.} \tag{19}$$

Note that this limit concerns the behavior of the match length at a fixed sequence index i as the window length tends to infinity. A similar result holds for the average match length over the entire sequence.

Theorem 6 For stationary ergodic sources with finite memory,

$$\frac{\mathbf{E}\{L_i(n)\}}{\log n} = \frac{1}{H} + \frac{O(1)}{\log n} \tag{20}$$

for any i and a fixed n. $O(1)$ represents some constant that does not grow with n. A very similar result holds for sources satisfying the Doeblin condition. Thus, for almost any experimental data, this theorem states that the expected match length is roughly inversely proportional to the entropy, with a vanishing constant, $O(1)/\log n$. Replacing the expectation in Eq. (20) with the sample mean over the observed sequence produces the sliding window match length (SWML) entropy estimator.

Definition 6 (SWML entropy estimator). The SWML estimator for the entropy of a stationary ergodic process satisfying the Doeblin condition is

$$\hat{H}_{\mathrm{SW}}(n) = \ell \left[\sum_{i=3}^{\ell} \frac{L_i(n_i')}{\log n_i'} \right]^{-1}, \tag{21}$$

where

$$n_i' = \min\{i-1, n\}. \tag{22}$$

where n is specified window size and n_i' is the effective window size. This estimator is asymptotically unbiased as $n \to \infty$ with probability one. Informally, n_i' is introduced to alleviate transient issues at the start of the data sequence when the index $i < n$.

The SWML estimator has several desirable properties. First, the SWML estimator is applicable to a broader class of sources than the model-based estimators, i.e., Classes (2)–(5) of the list in Sec. I B 2, compared to only Classes (4) and (5) for the Markov model-based estimators, and only Class (5) for the i.i.d. model estimator. Second, the SWML is robust to considerable departures from the assumptions of Eq. (21). When these assumptions are violated, $\hat{H}_{\mathrm{SW}}$ is still a valid upper bound on the entropy. Third, $\hat{H}_{\mathrm{SW}}$ converges rapidly in sequence length ℓ, producing a good entropy estimate from a relatively short data sequence. Fourth, and finally, varying the window size n provides a means to trade off between bias and adaptability to nonstationary sources, as Sec. I B 7 will address.

Despite a superficial similarity, there is no theoretical link between the SWML window size and the Markov model order, or between the SWML estimator and the Markov model. Although the SWML uses a fixed length portion of the immediately previous output, it does not estimate the empirical distributions of Eqs. (16) and (17) for the Markov estimator, Eq. (15).

The ability of the i.i.d. and Markov models to represent the structure of a song can be evaluated by applying each of the three entropy estimators to an observed sequence of units.

Property 3 When comparing two models for a source, the model which most accurately reflects the structural constraints of the source will produce a lower entropy estimate for the source's output.

Consequently, one can use the entropy estimates produced by two models to assess which of the models better fits the structural constraints embodied by a sample sequence. The model producing the lower entropy estimate is a better fit to the constraints of the source. An important caveat to this claim is that care must be taken to ensure that the Markov model estimates are accurate, and not artificially low, as discussed in Sec. I B 5.

Note that the SWML estimator [Eq. (21)] has the least restrictive assumptions of the three methods, and is therefore appropriate for the widest class of sources. Ideally, in the absence of bias, the SWML estimator should yield the smallest entropy estimate of the three estimators. The exception to this statement is if the source is actually an i.i.d. or Markov source. In this case, the SWML entropy estimate should roughly equal the i.i.d. or Markov entropy estimate. If the SWML and i.i.d. or Markov entropy estimates are roughly equal, this indicates that the i.i.d. or Markov model embodies all of the structural constraints apparent in the sample sequence analyzed.

5. Limitations on Markov models

Numerous animal communication studies have employed Markov models, or transition probability analysis (*c.f.*, Gentner and Hulse, 1998; MacKay, 1972; McCowan *et al.*, 1999; Slater, 1973). The empirical application of Markov models has clear limitations on the model order as a function of observed sample length. This section reviews these shortcomings to illustrate the advantages of nonparametric entropy estimation techniques, such as the SWML estimator.

In theory, increasing the order of an empirical Markov model improves its ability to approximate an unknown source. In practice, the finite available observation length ℓ places a limit on the model order. The sequence length required to obtain a reliable estimate of the PMF increases exponentially in model order. Entropy estimates are sensitive to inaccuracies in the PMF. Markov models with improperly high orders will likely not see all of the transitions that the true source produces, or will see them with an incorrect frequency. Such a model is said to *overfit* the data, and generally produces a deceptively low and inaccurate entropy estimate. This is particularly perilous when these erroneously low entropy estimates are interpreted using Property 3 of Sec. I B 4. The low entropy estimates may lead one to conclude incorrectly that the model is a better fit than it actually is.

How long must a sample sequence be to estimate entropy reliably from a Markov model? The data must be sufficiently long to obtain accurate estimates of all the probabilities in Eqs. (12), (16), and (17), or their higher-order analogs. A kth-order Markov model contains $|\mathcal{A}|^{k+1}$ parameters, which are the individual transition probabilities. Consequently, the data observed must exceed $|\mathcal{A}|^{k+1}$ symbols, and preferably $|\mathcal{A}|^{k+2}$. According to Theorem 4, a Markov model of order k and entropy H typically produces roughly $2^{H(k+1)}$ blocks of length $(k+1)$. If the model order k is large enough such that $\ell - k < 2^{H(k+1)}$, some typical blocks do not occur, and therefore the empirical distribution and the entropy estimate are fatally flawed. (Marton and Shields, 1994, 1996). The Markov order k as a function of data length ℓ must grow more slowly than $c\log(\ell)/H$ for the Markov entropy estimate to converge to the true entropy as ℓ goes to infinity (Marton and Shields, 1994, 1996). These results imply that $\ell \geqslant 2^{H(k+1)}$ is a necessary, though not sufficient, condition for accurate entropy estimation. Lacking prior information about the source, we must assume the most conservative case that $H = H_{\max} = \log|\mathcal{A}|$ and $\ell \geqslant 2^{(k+1)\log|\mathcal{A}|}$, or $|\mathcal{A}|^{(k+1)}$. This length is an absolute minimum. Preferably, the data length should satisfy $\ell \geqslant |\mathcal{A}|^{(k+2)}$, to provide accurate entropy estimates for most sources. The $\ell \geqslant |\mathcal{A}|^{(k+1)}$ requirement may be insufficient unless there is substantial prior evidence that the source's PMF precludes many of the $|\mathcal{A}|^{k+1}$ possible transitions.

The sample length required by these conditions is often onerous. For example, in typical humpback song with $|\mathcal{A}| \approx 20$, even a first-order Markov model requires at least 400–8000 units of song, and a second-order model requires at least 8000–160,000 units. At an average singing rate of 2.5 s/unit, the second-order Markov model requires roughly 5.5 to 110 h of uninterrupted recordings of a single singing whale. For this reason, we limited our Markov model analysis to first order.[4]

The i.i.d. and Markov estimators exhibit a negative bias even when the sample length requirements are satisfied. Section I B 6 presents results on correcting this bias in the entropy estimates. This correction must not be confused with the underestimation of entropy that occurs when Markov models overfit data sequences of insufficient length. Neither bias correction nor bootstrapping approaches can overcome the difficulties of overfitting. Using Property 3 to interpret the low entropy estimates produced by overfitting Markov models results in incorrect conclusions about the appropriateness of higher-order Markov models. Consequently, empirical Markov models have limited viability in entropy estimation. The combined difficulties of the risk of misinterpreting the entropy estimates of overfitted models and the requirement of unrealistically long data sequences argue strongly against the higher-order Markov model analysis of empirical data. One exception to this conclusion involves those rare instances where the source is known *a priori* to be a Markov source, and sufficient data are available.

In light of these strongly worded cautions, how did Shannon (1951) successfully estimate the entropy of printed English? Shannon's insight was to avoid estimating the joint probabilities of the sample data set. Instead, he estimated the error probabilities of human subjects predicting the next letter in an English text from the previous letters. Shannon then used the error probablities to estimate upper and lower bounds on the entropy of English. Subsequent experiments modified the guessing experiment to a gambling game to obtain improved entropy estimates for English (Cover and King, 1978; Miller, 1954; Ch. 6 of Cover and Thomas, 1991; Levitin and Reingold, 1994). Human-based techniques (Cover and King, 1978; Shanon, 1951) produce lower entropy estimates than the SWML estimator (Kontoyiannis, 1997) applied to the same text. Humans thus find less uncertainty in the text than mathematical algorithms. Similarly, it

is possible that humpback whales experience less uncertainty about the sequence of units in a song than the SWML estimator indicates.

Shannon's experiment is often interpreted to support the argument that English can be statistically modeled arbitrarily well by increasing the order of a Markov model. This view was challenged by Chomsky (1956) and Miller and Chomsky (1963). They concluded that: (1) Finite-state Markov models are incapable of representing the recursive hierarchical structures of English grammar; (2) successively higher-order Markov model approximations to English do not converge to true English grammar; and (3) the number of parameters required by higher-order Markov models to reflect grammatical constraints would be immense even if they did model English accurately.

6. Statistical properties of entropy estimators

Interpreting the entropy estimates discussed in Secs. I B 3 and I B 4 requires an understanding of the estimators' statistical properties, such as their bias[5] and confidence bounds. This section assumes that the data sequences are sufficiently long for the model-based methods, precluding the degraded entropy estimates discussed in the previous section. Even when the observed data sequence is sufficiently long, the entropy estimators are negatively biased due to Jensen's Inequality [*c.f.*, Eq. (9) of Wyner *et al.*, 1998; Ch. 2 of Cover and Thomas, 1991].

Theorem 7 (Basharin, 1959). The bias of the i.i.d. estimator for a sample sequence x_1^ℓ from an i.i.d. source is

$$\mathbf{E}\{\hat{H}_0\} - H = -\frac{|\mathcal{A}| - 1}{2\ell} \log e + \frac{O(1)}{\ell^2}. \tag{23}$$

If ℓ is sufficiently large, the term $O(1)/\ell^2$ may be neglected to obtain a bias-corrected estimate.

Definition 7 (bias-corrected i.i.d. estimator).

$$H_0' = \hat{H}_0 + 0.72\frac{|\hat{\mathcal{A}}| - 1}{\ell}. \tag{24}$$

Similarly, the first-order Markov model entropy estimator can be bias corrected as:

Definition 8 (bias-corrected first-order Markov estimator).

$$\hat{H}_1' = \hat{H}_1 + 0.72\frac{|\hat{\mathcal{D}}| - |\hat{\mathcal{A}}|}{\ell}, \tag{25}$$

where the number of all observed units is $|\hat{\mathcal{A}}|$, with $\hat{\mathcal{A}}$ as defined in Eq. (13), and the number of all observed digram transitions is $|\hat{\mathcal{D}}|$. If $|\hat{\mathcal{D}}|$ is less than the true value, the negative bias of $\hat{H}_1$ will be only partially corrected, and there will be some residual bias. Basharin (1959) also provides an expression for the variance of $\hat{H}_0$, but this requires knowledge of the source PMF, so is not applicable to empirical data analysis.

The bias of the SWML estimator cannot be analytically formulated like the biases of the model-based estimators, although it is known to be a positive bias which is $O(1)/\log n$. Thus, the SWML estimator's bias vanishes slowly with increasing window size n. Correcting the bias of the SWML estimator with a bootstrap technique is possible if replicate sample sequences can be generated having the same statistics as the *real* source. For example, Wyner *et al.* (1998) employed a Markov-based bootstrapping bias correction for the SWML estimator when estimating the entropy of DNA. However, in most animal communication research, including the present study, there is no *a priori* justification for assuming that the source follows a Markov model, and therefore it is inappropriate to correct the positive bias of the SWML entropy estimate with Wyner *et al.*'s (1998) technique.

7. Stationarity revisited

As Slater (1973) and MacKay (1972) argued, the processes producing animal vocalizations are not likely to be stationary, which limits the applicability of Markov models. Determining whether an unknown source is nonstationary is not always simple. The autocorrelation technique introduced in Sec. II C provides an indication of the stability of the time-local sequence statistics. If the source is stationary, the correlation values computed for the same lag in different segments of the sequence should be very similar. If large deviations are observed in the correlation estimates, the source is likely to be nonstationary. The statistical properties of many nonstationary sources change so slowly that a short subsequence of the observed data can be regarded as taken from a stationary source; such a source is called *locally stationary*. The i.i.d. and Markov models are strictly limited to stationary sources, and thus produce inaccurate entropy estimates for locally stationary sources. In contrast, the coding bound argument [Eqs. (2), (3), and (5)] justifies the use of the SWML estimator as a practical estimator for nonstationary processes, if the estimate is understood as an upper bound on the source entropy. It is desirable for this upper bound to be as tight as possible. To this end, the SWML estimator employs the following heuristic adaptation for the window size n.

Consider the effect of increasing or decreasing the SWML window size for a globally nonstationary but locally stationary source. Decreasing the window size makes the subsequence x_{i-n}^{i-1} within the window more likely to follow the same statistics as the matching sequence going forward from x_i. Unfortunately, decreasing the window size also increases the positive bias of $\hat{H}_{\mathrm{SW}}$, as can be seen from Eq. (20). Increasing the window size will reduce this bias until the window size n exceeds the limits of local stationarity, and the windowed sequence x_{i-n}^{i-1} no longer has the same statistics as the matching sequence x_i^∞. When the statistics of x_{i-n}^{i-1} and x_i^∞ differ, the observed match lengths are less than they would be for a stationary source of the same entropy. Referring to Eq. (21) reveals that decreasing the match length $L_i(n)$ increases $\hat{H}_{\mathrm{SW}}$. Thus, for a locally stationary source, choosing n to be either too small or too large will increase the positive bias of $\hat{H}_{\mathrm{SW}}$. To balance these competing demands, $\hat{H}_{\mathrm{SW}}$ is chosen to be the smallest entropy estimate obtained over a set of allowable window lengths $\mathcal{I}$, i.e.,

$$\hat{H}_{\mathrm{SW}} = \min_{n \in \mathcal{I}} \hat{H}_{\mathrm{SW}}(n). \tag{26}$$

Thus, the SWML estimator can operate even when the source is not stationary, at the small cost of an increased positive bias.

II. METHODS

The 16 songs analyzed were recorded off the coast of Hawaii from winter 1976 to spring 1978. These recordings of solo whales singing were originally analyzed in Payne *et al.* (1983). Each tape recording began at an arbitrary time during the sequence of units, and finished when the tape ended or the signal faded away. The longest recorded segments in this data set were roughly 45 min, and contained 1000 to 1200 units. Songs shorter than 300 units were rejected because they are insufficiently long for accurate entropy estimates.

The analysis method consists of two steps. First, the audio recordings of the humpback whale songs are converted into a sequence of symbols, where each symbol represents a distinct type of unit. Second, the entropy and correlation properties of the symbol sequences are estimated. Interpreting the entropy estimates' implications for the song structure requires two hypothesis tests. The confidence intervals used for these tests are presented in this section.

A. Classification

Classification necessarily causes an argument: "…there is no guarantee that we will draw the perceptual boundaries in the same place as our study animals" (Tyack, 1998). Ideally, "any human- or computer-generated categorization of vocalizations will need to be validated by testing with the species producing the calls" (Tyack, 1998). Such validating experiments are extremely difficult in the case of humpback songs, because all of the study animals are wild and very large, the units change over time, and the song structure also evolves. Furthermore, the perceptual boundaries of a specific animal may vary over time or depend on the behavioral or communicative context. These boundaries may also vary among the study population. Therefore, one cannot *a priori* assume that there is only one correct classification. The results presented in Sec. III are reassuringly robust to variations between different classifications.

Janik (1999) compared classification methods using human observers with computer-based methods. He remarked that the major disadvantages of human observer classification are bias and lack of reproducibility, but these can be mitigated and minimized by using several observers. Additionally, he noted that the design of automated classifiers must select both appropriate feature parameters and appropriate weightings for these features. In the current study, two human observers and a computer-based classifier were used to supplement each other's shortcomings, and each classification result was analyzed separately. The entire spectrogram of each unit was the input to an unsupervised neural network classifier to avoid the issue of parameter selection.

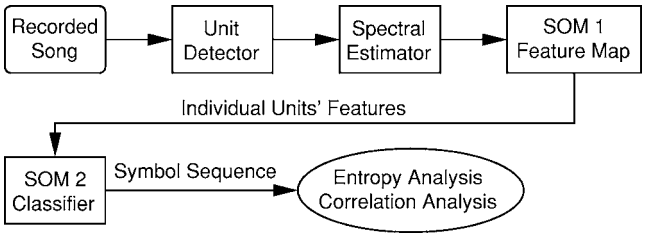


FIG. 2. Block diagram of the automated classifier analysis system for the humpback song.

In an analogous problem, human speech recognition systems, performance improves when the choice of the frequency scale is based on known properties of auditory perception (Jankowski *et al.*, 1995). Although little is specifically known about the hearing of humpback whales, the mammalian ear in general can be modeled as a bank of constant-Q bandpass filters called the cochlea filter (Pitton *et al.*, 1996). Also, humpback whale hearing appears to be low-frequency oriented, roughly matching their vocalization range of 20 Hz to a few kilohertz (Ketten, 1997). For these reasons, constant-Q (logarithmic) frequency-scale spectrograms form the input to the classifiers employed, rather than the more conventional constant bandwidth (linear) frequency-scale spectrogram.

The units in each song were classified using two methods: (1) Manual classification by two human individuals; and (2) automatic classification. Each of the roughly 20–45 min long analog recordings was digitized into a computer using a sampling frequency of 8000 Hz and 16-bit quantization. This digitized signal was used to produce the spectrograms used by both the automated classifier and the humans.

The two human-classified versions were independently produced by the individuals listening to the recorded song with the aid of printed spectrograms. The human classifiers listened to the song as many times as desired, and classified the units by writing the corresponding symbols onto the printed spectrograms. The spectrograms were computed using a polyphase filterbank and covered 30 Hz to 2700 Hz in 80 logarithmically spaced frequency bins and 200 time bins each representing 64 ms. The symbol sequences were typed into a computer file and doublechecked before further processing. The human classifiers were instructed to attach greater importance to the aural qualities of the sound than the printed spectrogram, whereas the automated classification was solely based on the spectrogram. The entropy estimators were applied separately to the symbol sequences produced by each classifier

The rest of this section describes the details of the automated classification. Figure 2 illustrates the automated classification and analysis system in block diagram form. An unpublished nonlinear method (Suzuki, in preparation) detected the signal of the individual units within the recording. The polyphase filterbank computed the spectrograms of these individual units, along with a short sample of the background noise that preceded and followed the unit.

Prior to classification, each unit's spectrogram was preprocessed as follows: (1) The spectrogram was centered within the 12.8 s time window,[6] with any time bins before or

after the signal zero filled; (2) all bins were normalized by the maximum bin; (3) all bins below a threshold were set to zero; (4) soft clipping was applied to all bins above a threshold; and (5) the entire spectrogram was normalized for unit energy. This preprocessing improved the performance of the classification over the range of signal amplitudes and background noise during the song. The processed spectrograms of the units were the inputs to the automated classifier consisting of a sequence of two self-organizing maps (SOMs) (Kohonen, 2001). This classifier assigned a letter to each unit's individual spectrogram, where (ideally) each distinct letter represented a group of similar spectrograms. The sequences of units in the song recording were thus mapped to a symbol sequence. The work of Walker *et al.* (1996) strongly influenced the design of our classifier.

The SOM classifier does not require an *a priori* decision regarding the alphabet size $|\mathcal{A}|$, but only sets the maximum allowable $|\mathcal{A}|$. This is a distinct advantage over classifiers such as the *k*-means clustering algorithm (Linde *et al.*, 1980) or a learning vector quantizer (LVQ) (Kohonen, 2001; Chap. 10 of Duda *et al.*, 2001), since it avoids a major subjective decision about the exact number of classes present in the data. This advantage led us to use the SOM for our classifier in spite of Kohonen's advice advocating LVQs over SOMs for classification and decision tasks (Kohonen, 2001).

The first-stage SOM in the classifier groups units with similar spectrograms, implementing a common SOM topological feature map (Haykin, 1999; Kohonen, 2001). This stage is organized as a relatively large two-dimensional lattice with a large slowly decaying topological neighborhood. The output of this first stage is a matrix of similarity values between the input spectrogram and each neuron's trained synaptic weight. The second-stage SOM takes the similarity matrix for a unit as its input and produces a discrete symbol in $\mathcal{A}$ as its output. The second-stage SOM is organized as a one-dimensional lattice whose length is determined by the maximum allowed value for the alphabet size $|\mathcal{A}|$. This SOM has a small, rapidly decaying neighborhood suitable for classification.

The classifier was independently trained for each song processed. Both SOMs were trained using spectrograms of randomly chosen units from the song. For the first-stage SOM, each unit was equally likely to be chosen as the input. After the first-stage SOM was trained and the weights were fixed, the second-stage SOM was trained twice, using the similarity matrices output by the first SOM for randomly chosen units. For the first training run of the second-stage SOM, each unit spectrogram was equiprobable. For the second training run of the second-stage SOM, each output letter was chosen equiprobably, and then an individual spectrogram for an instance of that letter was chosen equiprobably from all spectrograms classified as that letter during the first run. Therefore, during the second training run, spectrograms belonging to infrequently appearing units were chosen more frequently than the spectrogram of common units, but all possible classifier output letters received approximately the same number of training iterations.

Both SOMs were trained with the standard adaptation techniques found in Kohonen (2001) and Haykin (1999), with typical parameters given in Table I using the notation of Chapter 9 of Haykin (1996). The topological neighborhood function was a Gaussian function as suggested in Haykin (1999). The only modification from the standard SOM neuron update algorithm was that, for the first-stage SOM, the spectrogram matrix could be shifted by up to Δt bins in time and Δf bins in frequency to determine which neuron's weights were the most similar to the current spectrogram. For the first 2000 iterations of training, the neurons' weights were then updated using the unshifted original spectrogram. After 2000 iterations, the neurons' weights were updated using the shifted spectrogram which gave the best fit to the optimal neuron. For the purpose of these shifts, both the spectrogram and the neuron weights were considered to be zero outside their defined index range. Table I also includes typical values for Δt and Δf.

TABLE I. Neural network parameters. These are the typical parameters used for the two stages of self-organizing maps (SOMs) in the automatic classification system.[a]

Parameter	First stage	Second stage
Map size	8×8	1×26
Input dimension	80×200	8×8
Iterations	40,000	80,000
σ_0	8	7
τ_1	850	8
η_0	0.1	0.03
τ_2	1000	800
Δt	10 bins	⋯
Δf	6 bins	⋯

[a]The notation for the SOM parameters in this table follows Chapter 9 of Haykin (1999).

Once the SOMs for both stages of the classifier converged, the spectrograms of the units were processed again to produce the automated classifier's symbol sequence for that song.

B. Information theory analysis

In order to determine whether the i.i.d. and Markov models accurately represented the structure embodied by each symbolized humpback song, the model-based entropy estimates and the SWML entropy estimate were used to test two hypotheses. The first hypothesis is that $H_0 \leqslant H_1$, with the alternate hypothesis being $H_0 > H_1$. The second hypothesis is that $H_1 \leqslant H_{\mathrm{SW}}$, with the alternate hypothesis being $H_1 > H_{\mathrm{SW}}$. Decisions of these tests are interpreted with Property 3 of Sec. I B 4.

The two hypotheses are tested for the significance level of 0.05 using the following bootstrap technique. For the first hypothesis ($H_0 \leqslant H_1$), 1000 independent bootstrap sequences are generated for each song using an i.i.d. source whose PMF is the empirically observed distribution for the song. Each of these bootstrap sequences has the same length as the original song sequence. The bootstrap sequences are used to obtain 1000 first-order Markov entropy estimates. The 50th lowest of these entropy estimates is chosen as the bound on the one-tailed 0.95 confidence interval for the source entropy measured with $\hat{H}_1$, under the assumption that the actual

TABLE II. Entropy estimates from 16 humpback whale songs using the three estimators described in Secs. I B 3 and I B 4. All entropy estimates are in bits. The i.i.d. and Markov values are bias corrected as described in Sec. I B 6. The number in parentheses next to the $\hat{H}_{\mathrm{SW}}$ indicates the window size n selected as described in Sec. I B 7. $\hat{H}_{\mathrm{SW}}$ is consistently less than both $\hat{H}_0$ and $\hat{H}_1$, demonstrating that the Markov and i.i.d. models fail to capture the full structure of the songs. The strength of the song structure is also apparent in the high redundancy values for all songs.

Date (Tape No.)	Length (units)	$\lvert\hat{\mathcal{A}}\rvert$ (units)	$\log\lvert\hat{\mathcal{A}}\rvert$ $H_{\max}$	i.i.d. $\hat{H}_0$	Markov $\hat{H}_1$	SWML $\hat{H}_{\mathrm{SW}}(n)$	Redundancy $\hat{\rho}$
26 Dec 1976 (1A)	840	21	4.39	4.02	0.84	0.58(10)	0.87
04 Jan 1977 (2)	1103	17	4.09	3.79	1.13	0.64(10)	0.84
01 Feb 1977 (1-1)	972	27	4.75	4.13	0.99	0.47(10)	0.90
04 Mar 1977 (1A-2)	978	25	4.64	3.93	0.75	0.33(11)	0.93
10 Mar 1977 (1A-1)	967	27	4.75	4.22	1.17	0.68(11)	0.86
10 Mar 1977 (1A-2)	577	26	4.70	4.03	0.99	0.57(11)	0.88
12 Apr 1977 (4A)	805	33	5.04	4.54	1.10	0.56(10)	0.89
17 May 1977	1021	17	4.09	3.61	1.12	0.36(10)	0.91
01 Feb 1978 (1A)	1101	26	4.70	4.38	1.10	0.56(11)	0.88
05 Feb 1978 (1A)	976	18	4.17	3.77	0.81	0.51(13)	0.88
05 Feb 1978 (1B)	959	18	4.17	3.88	0.83	0.58(11)	0.85
05 Feb 1978 (2A-1)	380	23	4.52	4.37	0.98	0.52(13)	0.88
05 Feb 1978 (2A-2)	488	25	4.64	4.36	1.05	0.64(13)	0.86
05 Feb 1978 (2B-1)	438	22	4.46	4.26	0.94	0.67(15)	0.85
05 Feb 1978 (3A-2)	436	21	4.39	4.37	1.01	0.60(11)	0.87
07 Feb 1978 (1B)	662	26	4.70	4.45	1.10	0.79(15)	0.83

humpback song is generated with an i.i.d. model. If the observed $\hat{H}_1$ for the song is below this value, the null hypothesis is rejected. A similar bootstrap procedure is used for the second hypothesis ($H_1 \leqslant H_{\mathrm{SW}}$), except that the empirical first-order Markov model is used to generate the 1000 replica sequences, and SWML estimator is used for the entropy estimation. The SWML entropy estimates were not constrained to use the same window size n in Eq. (26) which obtained the minimum H_{SW} for the song, but rather could range over the full set $\mathcal{I}$ for each replica sequence. The resulting confidence bound is for the source entropy measured with H_{SW} assuming that the actual humpback source is a first-order Markov source.

C. Correlation analysis

Rejecting both null hypotheses described above indicates that neither the i.i.d. nor the first-order Markov models are adequate to produce the structure of the observed humpback song. This does not rule out the possibility that the song was produced by a higher-order Markov model. As discussed in Section I B 5, entropy estimates for second- (and higher-) order Markov models will be unreliable for humpback songs with an alphabet of size $|\mathcal{A}| \approx 20$ and length $\ell \approx 300-1200$ units. Instead, correlation analysis can be used to reject the possibility of a higher-order Markov model.

The discrete sequence correlation of two symbol sequences x_1^{ℓ} and y_1^{m} at lag λ is defined by

$$r(x_1^{\ell}, y_1^{m}, \lambda) = \frac{\left|\{i : x_i = y_{i+\lambda}, \max(1, 1-\lambda) \leqslant i \leqslant \min(\ell, m-\lambda)\}\right|}{\min(\ell, m, \ell+\lambda, m-\lambda)} \tag{27}$$

for $|\lambda| < \min\{\ell, m\}$. The numerator represents the number of symbols of x_i and $y_{i+\lambda}$ that agree within the overlapping section. The denominator is the length of the overlapping section of the sequences after y has been shifted by λ. Note that this definition can also be used for autocorrelation if y is set equal to x.

A Markov model of any order is stationary, so the value of the short time autocorrelation $r(x_n^{n+L}, x_n^{\ell}, \lambda)$ should be independent of the value of n chosen, i.e., which segment of the sequence is used for the autocorrelation. In practice, the autocorrelation for a fixed value of λ will fluctuate slightly with n, but should be very similar if the process is stationary. If the estimates of the autocorrelation for different values of n are only similar over a range of lags $0 \leqslant \lambda \leqslant v$, but diverge for $\lambda > v$, this implies that the source is nonstationary, but may be considered locally stationary for windows of length v or less. This value of v determines the range of window sizes $\mathcal{I}$ used for the SWML estimator in Eq. (26).

To determine the stationarity of the humpback song, disjoint sections of longer songs was autocorrelated using two 151-unit sections from the song. Specifically, $r(x_{300}^{450}, x_{300}^{\ell}, \lambda)$ and $r(x_{600}^{750}, x_{600}^{\ell}, \lambda)$ were compared to see if the correlation function differed early and late in the recording.

In addition to this short-time autocorrelation analysis, global autocorrelations were computed using $r(x_1^{\ell}, x_1^{\ell}, \lambda)$ for the songs. These autocorrelations, considered as a function of λ, reveal the timescales of the dependencies and periodicities in the symbol sequence. Sequences which are periodic or quasi-periodic with period N will produce global autocorrelations with the same period. Comparing global autocorrelations from different songs provides information about how the periodicity changes between songs.

III. RESULTS

A. Entropy analysis

Table II presents the three entropy estimates for each song, converted to symbol sequences by one of the human

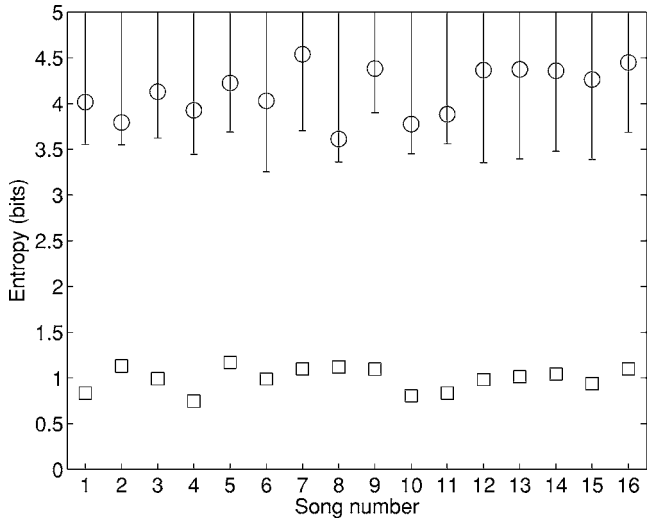


FIG. 3. The values of $\hat{H}_0$ (circles) and $\hat{H}_1$ (squares) plotted for each song in the same order as Table II. The interval plotted for each song indicates the lower limit of the one-tailed 0.95 confidence region for $\hat{H}_1$ under the hypothesis that $H_0 \leqslant H_1$. For all songs, $\hat{H}_1$ lies below the bars, supporting the conclusion that $H_1 < H_0$ with significance $p < 0.05$. This indicates that the humpback songs contain a temporal structure that partially depends on the immediately previous unit within a song, and that the i.i.d. model failed to capture the structures embodied by the humpback songs.

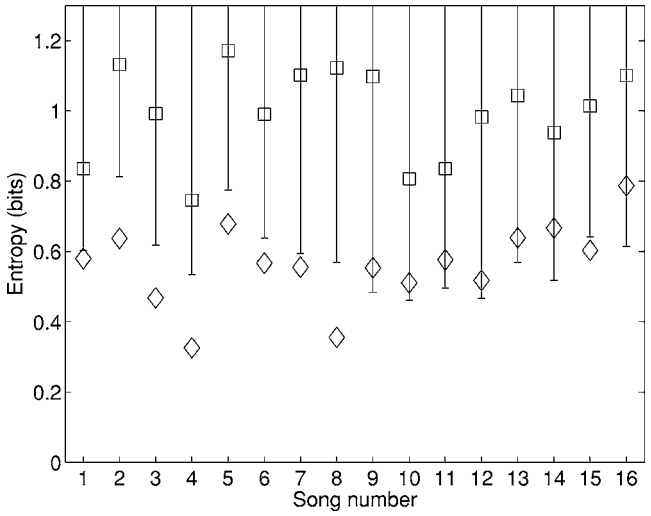


FIG. 4. The values of $\hat{H}_1$ (squares) and $\hat{H}_{\mathrm{SW}}$ (diamonds) plotted for each song in the same order as Table II. The interval plotted for each song indicates the lower limit of the one-tailed 0.95 confidence region for $\hat{H}_{\mathrm{SW}}$ under the hypothesis that $H_1 \leqslant H_{\mathrm{SW}}$. For 9 of the 16 songs, $\hat{H}_{\mathrm{SW}}$ lies below the interval, supporting the conclusion that $H_{\mathrm{SW}} < H_1$ with significance $p < 0.05$. This demonstrates that the Markov model failed to capture all of the structure embodied by the majority of the humpback songs we analyzed, and that the humpback songs contain temporal structure spanning over the range beyond immediately adjacent units.

classifiers. The values given for $\hat{H}_0$ and $\hat{H}_1$ are the bias-corrected values, and the values $\hat{H}_{\mathrm{SW}}$ use window sizes in the range $\mathcal{I}=[10,40]$. The number in parentheses following $\hat{H}_{\mathrm{SW}}$ is the window size yielding this minimum value in Eq. (26). Table II also includes the estimated redundancy $\hat{\rho}$ for each source, using $\hat{H}_{\mathrm{SW}}$ for the entropy H in Eq. (6). The high value of $\hat{\rho}$ for each song indicates that the song structure is strongly constrained, resulting in an entropy much less than $H_{\max}$ for the observed alphabet size $|\mathcal{A}|$.

Figure 3 presents the i.i.d. (circles) and Markov (squares) entropy estimates for individual songs with one-tailed 0.95 confidence bounds for the null hypothesis, $H_0 \leqslant H_1$. Since all $\hat{H}_1$ lie below the confidence bound, we reject the null hypothesis ($p \leqslant 0.05$) and conclude that $H_1 < H_0$. This means that the i.i.d. model failed to capture the structure embedded in all humpback songs analyzed, and also that the humpback songs contain a temporal structure that at least partially depends on the immediately previous unit within a song.

Figure 4 presents Markov (squares) and SWML (diamonds) entropy estimates for individual songs with one-tailed 0.95 confidence bound for the null hypothesis, $H_1 \leqslant H_{\mathrm{SW}}$. 9 out of 16 songs had $\hat{H}_{\mathrm{SW}}$ below this confidence bound, and for these songs we conclude that $H_{\mathrm{SW}} < H_1$ ($p \leqslant 0.05$). For the seven other songs, the test did not conclude that the $\hat{H}_1$ and $\hat{H}_{\mathrm{SW}}$ were sufficiently different at this significance level. We offer two possible explanations to account for this result. All of these seven songs were recorded in a one week period in early February 1978. Since they are likely to be quite similar, those data points may not be statistically independent. Considering only one song in any given month, the second hypothesis test concludes $H_{\mathrm{SW}} < H_1$ for six out of seven months. Furthermore, we note that $\hat{H}_1$ may have residual negative bias (Sec. I B 6) if $|\mathcal{D}|$ was underestimated, and $\hat{H}_{\mathrm{SW}}$ has positive bias (Sec. I B 6). Both sources of bias work to favor the null hypothesis by requiring stronger statistical evidence to reject the null hypothesis, but we disregarded these residual bias considerations in our analysis. Therefore, it is expected that the true test results may be more significant than the results presented in Fig. 4, although we have no way to demonstrate this. Although the result is less clearcut than the one shown in Fig. 3, the conclusion from Fig. 4 is that the Markov model failed to cap-

TABLE III. Automatic versus manual classification. The gaps between the different entropy estimates for each song persist regardless of the classifier used. Consequently, the conclusions about the song structure are robust to the classification method chosen. All entropy estimates are in bits.

		4 Jan 1977			17 May 1977		
		Auto	Human 1	Human 2	Auto	Human 1	Human 2
i.i.d. model	$\hat{H}_0$	2.84	3.71	3.79	4.38	3.61	3.61
Markov model ($k=1$)	$\hat{H}_1$	2.48	1.37	1.13	2.69	1.26	1.12
Sliding window	$\hat{H}_{\mathrm{SW}}(n)$	1.95(10)	0.89(10)	0.64(10)	1.59(18)	0.47(10)	0.36(10)

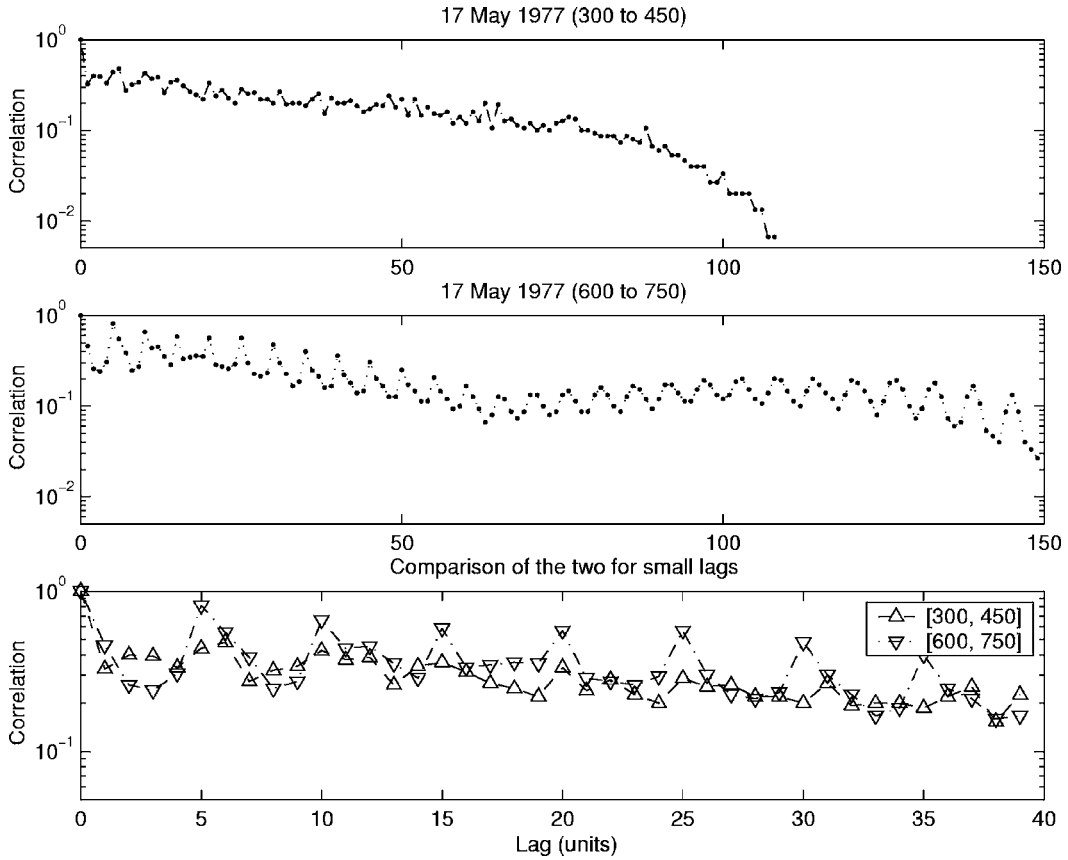


FIG. 5. Two short time autocorrelation estimates for a typical song. The top two panels plot the correlations for units in the ranges [300, 450] and [600, 750], respectively. Note that the correlation functions diverge substantially for lags greater than 100 units, indicating that the song is nonstationary. The third panel shows both correlations on the same axes for small lags. The values generally agree closely to each other with lags smaller than 40, indicating that the song may be considered locally stationary over this range. Both correlations have a strong oscillation with a period of about 6, corresponding to the typical phrase length of Payne *et al.* (1983).

ture all the structure embodied by the majority of songs we analyzed, and also that the humpback songs contain a temporal structure which spans beyond the immediately previous unit within the song.

To control for observer bias, the songs were classified by two humans and the SOM classifier described in Sec. II A. The human classifiers disagree on roughly 5% of the units for these songs, suggesting that there is no clear correct classification for all of the units. The automatic classifier disagrees with the human observers for roughly 10% of the units. Table III contrasts the entropy estimates obtained using each of the three classifiers on two songs. The changes in the entropy estimates between classifiers for a given model reflect the disagreements about the unit classifications. The entropy estimates vary for the different classifiers, but all three estimators exhibit substantial gaps between $\hat{H}_0$ and $\hat{H}_1$ and between $\hat{H}_1$ and $\hat{H}_{\mathrm{SW}}$. These gaps between the entropy estimates are also robust to variations in the SOM parameters. Therefore, the fundamental conclusions above about the songs' structure using Property 3 of Sec. I B 4 are robust to the choice of classifier. We speculate that the strong structural constraints of the songs prevent the perturbations in unit classifications from closing the gaps between the entropy estimates.

B. Correlation analysis

Local autocorrelation functions were computed to test the stationarity of the songs. The results for May 17, 1977 are shown in Fig. 5. The top panel plots $r(x_{300}^{450}, x_{300}^{1021}, \lambda)$ as a function of the lag λ, while the middle panel plots $r(x_{600}^{750}, x_{600}^{1021}, \lambda)$ for the same song. Both plots have logarithmic vertical axes. Comparing the top two panels of Fig. 5 reveals that the correlation curves differ substantially for lags of 100 units or more. This difference between the curves demonstrates that the song statistics are nonstationary. Any irreducible empirical Markov model has stationary statistics, so consequently no such Markov model with an order less than the maximum possible length of the song can capture the structure of the songs. Lastly, note that the oscillations with a period of 6–8 units in the autocorrelation function indicate that there are repetitions in the song with that period.

A possible objection arises to the use of $\hat{H}_{\mathrm{SW}}$ for the humpback songs, since the songs do not have stationary statistics and Eq. (21) assumes a stationary source. We offer two rationales in response to these points. First, even when the assumptions of the SWML entropy estimator are violated, the estimator produces an upper bound on the entropy (Kontoyiannis *et al.*, 1998), so the true entropy is expected to be

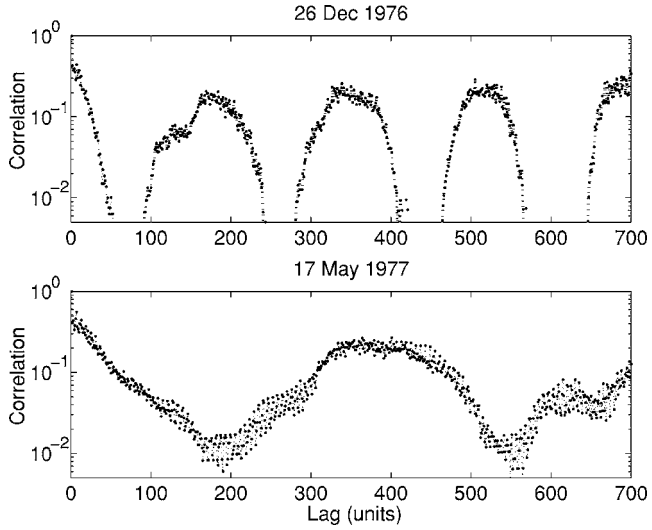


FIG. 6. Global autocorrelation estimates for two songs from the 1976–1977 season. Note that the December song has a period of about 180 units, and the May song of about 400 units. These periods are consistent with the song durations reported in Payne *et al.* (1983). Combining this information with the shorter period shown in Fig. 5, these figures demonstrate that the sequence of units in the songs has multiple periods.

even lower than the values reported in Table II. Thus, the conclusions drawn from Property 3 are still valid. Second, over the window lengths used to obtain $\hat{H}_{\mathrm{SW}}$, the source may be considered locally stationary, as discussed in Sec. I B 7. To justify this assertion, the bottom panel of Fig. 5 enlarges and overlays the two autocorrelation values for $0 \leqslant \lambda \leqslant 40$. This close agreement indicates that the song source may be considered to be locally stationary over windows of 40 units or less. Autocorrelation analysis of other songs exhibited similar agreement over this range of lags. This agreement determined our choice of $\mathcal{I} \in [10,40]$ for the window lengths range in Eq. (26).

The global autocorrelation function $r(x_1^{\ell}, x_1^{\ell}, \lambda)$ shown in Fig. 6 for two songs exhibits a superposition of two oscillations. The shorter period low amplitude oscillations are those of a period of 6–8 units seen in Fig. 5. On the larger timescale used for Fig. 6, these short period oscillations are no longer clearly discernible, but appear as a vertical blur. The larger amplitude oscillation has a period of roughly 180 units for the Dec. 1976 song, and roughly 400 units for the May 1977 song. The other songs analyzed also exhibited a long and short period oscillation. These oscillations demonstrate that the songs possess constraints repeating on segments of 6–8 units in length and also segments hundreds of units in length.

IV. DISCUSSION

Payne *et al.* (1983) previously analyzed a superset of the data used in this study, using human observers of the spectrograms to tabulate the song duration data over 31 day periods. During their Period II of 1976–1977, which includes Dec. 1976, the average song duration was 7.5 min. In their Period VI of the same season, which includes May 1977, the average duration was 13 min. The ratio of their averaged song durations is 1.7. On the other hand, using our correlation analysis, the periods of the Dec. 1976 and May 1977 songs are found to be 180 and 400 units, respectively. The ratio is 2.2. Using an average of 2.5 s per unit, the durations of these songs are approximately 7.5 and 16.7 min. The close agreement between these ratios and durations suggests that the longer period oscillations correspond to their song duration. Additionally, the 6–8 unit oscillations observed in the correlation functions of our Fig. 5 correspond closely to the phrase lengths of 4–10 for most phrases indicated on the spectrograms in Figs. 3 and 4 of Payne *et al.* (1983) and Figs. 6, 8, and 9 of Payne and McVay (1971).

Intriguingly, the entropy and the period of the songs vary over the 1976–1977 singing season. The entropy largely peaks early in the season (Dec. and Jan.) and decreases through the season. Recall from Property 1 of Section I B 2 that a constrained and predictable source has a lower entropy than an unconstrained or unpredictable one. The entropy variations through the season provide an objective quantitative confirmation of several subjective observations made in Payne *et al.* (1983). They observed that "the whales were the most consistent in terms of which themes they included between late 1977 and early 1978… when the number of themes in each song was highest." The relatively low entropy value in May 1977 corroborates this observed consistency in theme selection.

Payne and McVay (1971) defined a transitional phrase to be a phrase occurring between two themes which combines features of both. Some transitional phrases mix units from two adjacent themes in a complex way. Fig. 17 of Payne *et al.* (1983) plots the mean percent of the song duration devoted to transitional phrases. For the 1976–77 season, Period II has the highest proportion of transitional phrases (5%), and the proportion monotonically decreases toward the end of the season. These complicated unusual phrases should increase the total entropy of the song, and the decrease in the proportion of transitional phrases is consistent with the trend found in this analysis of decreasing entropy estimates through the season. Similarly, Payne *et al.* (1983) observed "During the time when the song was least stable in terms of which component themes were present, it contained transitional phrases between all themes; but at the end of the 1976–77 season, when all themes were firmly established, there were no transitional phrases left…. In other words, the song was maximally compartmentalized, organized, and predictable." Again, this qualitative observation is consistent with the general trend of decreasing entropy toward May 1977.

Guinee and Payne (1988) suggested that songs with larger number of themes (longer songs) are often more redundant in phrase structures than shorter songs. Such redundant long songs would be expected to have lower source entropies than shorter songs. Phrased in terms of the entropy analysis, the song lengths determined by the global correlation analysis and the source entropy of the song should be negatively correlated. This would roughly conserve the total amount of information required to remember a song's structure. Our data generally confirm this for the seven independent song sessions analyzed from the 1976–1977 season. (Note that in Table II, there are two segments from the same

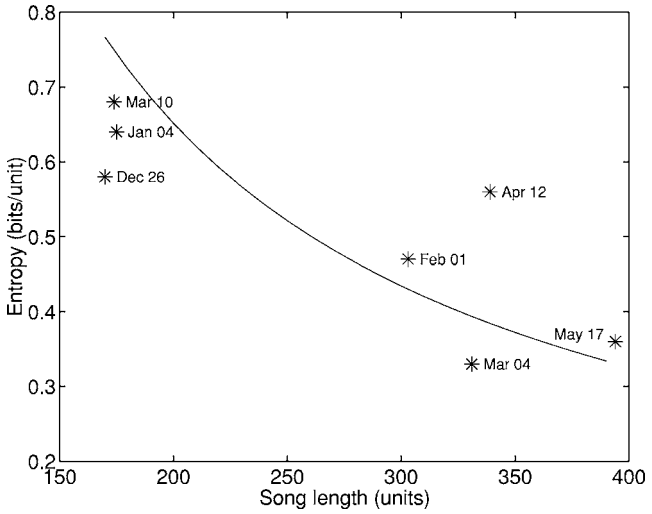


FIG. 7. This scatter plot represents the general trend of increasing song length and decreasing source entropy during the 1976–1977 season observed in seven songs. With the exception of one anomalous singer (Mar. 10), the period of the song increases from Dec. 1976 through May 1977. The source entropy generally decreases through the season, but not as steadily as the period increases. The solid line represents the best fit of a model in which the product of the period and the source entropy is constant for the season. This product, 130.3 bits/song, represents the average total information in the sequence of units in the song.

song session on Mar. 10, and thus these are not really independent observations.) Figure 7 plots the average period versus the source entropy for these seven song sessions. The average period was estimated from the global autocorrelation functions. The product of the song length and source entropy provides a measure of the average amount of total information in the song. One possible relation consistent with Guinee and Payne's (1988) observations is that this product, the bits per song, should be roughly constant through the season. The solid line in Fig. 7 indicates 130.3 bits per song, which is the average of the product of the period and $\hat{H}_{\mathrm{SW}}$ for the song sessions plotted in this figure. Although this relation may be an overly simplistic one, the curve loosely fits the data points, with clear outliers. These observations about the total information are made from a limited number of songs. A more diverse body of data is needed to reach general conclusions.

The results of this study have implications for larger questions comparing the structures of human and nonhuman animal acoustic communications. Some researchers have hypothesized that nonhuman animal acoustic communications lack recursive structures (Chomsky, 1988; Hauser *et al.*, 2002). A recent survey noted that "long-distance hierarchical relations are found in all natural languages, for which, at a minimum, a 'phrase structure grammar' is necessary," (Hauser *et al.*, 2002). The same authors observed that "the core recursive aspect of FLN [the Faculty of Language—Narrow Sense] currently appears to lack any analog in animal communication" and conjectured that "FLN—the computational mechanism of recursion—is recently evolved and unique to our species" (Hauser *et al.*, 2002). This hypothesis is supported by Fitch and Hauser's recent results demonstrating that cotton-top tamarins (*Saguinus oedipus*) are able to parse synthetic stimuli sequences generated by finite-state grammars, but not those generated by a simple recursively hierarchical (phrase structure) grammar (Fitch and Hauser, 2004). Researchers studying human languages describe the property of creating an infinite number of valid signals from a finite set of discrete units as "discrete infinity." Chomsky (1988), in discussing this property, conjectured that "Human language has the extremely unusual, possibly unique, property of discrete infinity" and that "language is based on an entirely different principle than any animal communication system."

The hierarchical structure proposed by Payne and McVay (1971) for humpback whale song challenges these conjectures on the uniquely human nature of long-distance hierarchical relations, and potentially on the uniquely human property of recursion and discrete infinity. Hierarchical grammars may be efficiently represented using recursion, although recursion is not necessarily implied by hierarchy. Additionally, the proposed hierarchical structure and continuous modification of the song raise the possibility that humpback whales can, in theory, create an infinite number of valid songs from the finite set of discrete units. The entropy comparisons in Sec. III A demonstrate that the humpback songs contain structures more complex than first-order Markov models. The lack of stationarity found in the short-term correlation properties of the songs in Sec. III B establishes that the songs cannot be modeled by a higher-order Markov model. The multiple periods on the order 6–8 and hundreds of units found in the autocorrelations are consistent with the hierarchical structure for the songs proposed by Payne and McVay (1971). These multiple periods in the autocorrelation functions illustrate that humpback songs "go beyond purely local structure," demonstrating the sort of "statistical regularities that are separated by an arbitrary number of words" which Hauser *et al.* (2002) asserted are a consequence of natural languages. A hierarchical grammar is a simple and efficient model to produce multiple periods in a song's structure. In light of the previously mentioned inability of some primates to parse hierarchically organized acoustic stimuli (Fitch and Hauser, 2004), this evidence of hierarchical structure in the songs of the evolutionarily more distant humpback whales is intriguing.

It is important to emphasize that we do not claim that humpback songs are a language in the sense recognized by linguists. Hauser *et al.* (2002) define the "conceptual-intentional" component of a language as the property that the different sentences produced as the words are rearranged within the grammatical structure "differ systematically in meaning." There is no evidence that humpback songs satisfy this linguistic requirement. Payne and McVay (1971) made no claims that the large potentially infinite variety of songs produced by the hierarchical ordering of the discrete units have distinct meanings. Tyack (1981) also speculated that each change in the song does not correspond to a change in what the song communicates to other whales. More recently, Payne (1995) explicitly asserted

> "As regards the possible language content of humpback whale songs one can say with certainty only that no one knows whether they contain anything at all that we would equate with language. At the

present state of knowledge, claims to the contrary are simply speculation."

For any one of the components of language enumerated by Hauser *et al.* (2002), it appears possible that there is some animal communication system which possesses that property while lacking one or more of the other linguistic components. Bonobo (*Pan paniscus*) lexigrams appear to satisfy the conceptual-intentional requirement of language (Savage-Rumbaugh *et al.*, 1986), but there are no clear data supporting hierarchical or recursive structure in these animals' communications. The results of the present study are consistent with a hierarchical and recursive structure for humpback song, but as noted above, there is no evidence for humpback song satisfying the conceptual-intentional property of language. This evidence supporting hierarchical structure in humpback song, taken in conjunction with recursion's known efficiency in representing hierarchical structures, raises questions about Hauser *et al.*'s (2002) conjecture that only humans employ recursion to structure their communication signals.

V. CONCLUSION

This paper presents an overview of modern information theory in the context of the analysis of animal communication systems, and then applies these techniques to humpback whale song. The SWML entropy estimator is proposed as a replacement for the i.i.d. and Markov model entropy estimators commonly employed to study animal communications. The SWML estimator is applicable to a much broader class of information sources than the model-based estimators, and its estimates provide valid upper bounds on the entropy for nonstationary sources. The SWML estimator also converges rapidly—an important characteristic for animal communication studies where it is often challenging to obtain long data sequences. In contrast, high-order Markov models risk underestimating the entropy and drawing erroneous conclusions about the source unless extremely large data sets are available.

The entropy and correlation properties are estimated for 16 humpback whale songs. The entropy estimates for both the i.i.d. and first-order Markov models exceed the SWML estimate, indicating that these models fail to capture the structure of the song. The results hold for songs transcribed by two humans and a computer program, controlling for any subjective bias. The SWML estimates indicate that the amount of information carried by the sequence of the units in the song is less than 1 bit per unit. Combining the SWML entropy estimate with the period estimate from the correlation analysis gives an average of roughly 130 bits per song. The correlation analysis demonstrates that songs are nonstationary, but may be considered locally stationary over segment lengths of roughly 40 units. This nonstationarity of the source means that no irreducible empirical Markov model can represent the song structure. The correlation data also indicate that the songs are periodic on two scales of approximately 6–8 and 200–400 units. Such a correlation structure is simply and efficiently produced by hierarchical models—consistent with the Payne and McVay (1971) grammar—but difficult to produce without hierarchy. The entropy and correlation results also provide quantitative confirmation of several observations made by Payne *et al.* (1983) about the song's evolution during the 1976–1977 season. The correlation data demonstrate that the songs possess strong long-distance dependencies of the sort discussed in Hauser *et al.* (2002) as a hallmark of phrase structure grammar. In closing, there is substantial quantitative evidence consistent with the sequence of units in the humpback songs being organized in a hierarchical structure, but equally strong evidence that this sequence is carrying relatively little information.

ACKNOWLEDGMENTS

The authors thank Roger Payne for allowing us to reanalyze the tapes originally made for Payne *et al.* (1983), and Ashley Walker for providing us with unpublished information and programs which inspired us to apply SOMs as classifiers for humpback songs. They also thank Hanspeter Herzel and Tecumseh Fitch for helpful discussions in evaluating and interpreting preliminary results in this study. Hanspeter Herzel also brought Basharin (1959) to our attention. Yoram Bresler and Andrew Solow encouraged us to evaluate the statistical properties of the entropy estimators. Phillip Pendergrass patiently classified hours of humpback song to make this study possible. Two of the authors (J.R.B. and R.S.) gratefully acknowledge the support of the NSF (Ocean Science Career Award No. 9733391). While preparing the final revision of the manuscript, one of the authors (R.S.) was an MIT Rosenblith Fellow and Howard Hughes Medical Institute Predoctoral Fellow. Another author (J.R.B.) received additional support from the Australian-American Fulbright Commission during the revision of the manuscript. ONR Grant No. N00014-00-1-0379 provided the computer facilities used for the entropy estimation. An author (P.L.T.) wishes to acknowledge support from ONR Grant No. N00014-97-1-1031. This is contribution No. 99-1102 from the University of Massachusetts Dartmouth School for Marine Science and Technology and No. 10093 from Woods Hole Oceanographic Institution.

APPENDIX A. PROGRESSION OF INFORMATION ENTROPY AND ITS ESTIMATION

Information entropy was first defined for stationary ergodic Markov sources and the entropy theorem was proved for i.i.d. sources by Shannon (1948). Khinchin (1953, 1957) reinforced Shannon's concepts with rigorous mathematical treatments and proved the entropy theorem for stationary ergodic Markov processes. McMillan (1953) extended the entropy theorem to stationary ergodic sources with convergence in probability. Breiman (1957, 1960) sharpened the convergence of the entropy theorem to convergence with probability one. McMillan (1953) named this theorem the *asymptotic equipartition property*. However, a more intuitive name, the entropy theorem, from Shields (1996), is used in this paper. The coding theorem and its converse for noiseless encoders were first proved by Khinchin (1953, 1957). While Shannon's theory is based entirely on probability theory, Kolmogorov (1965) proposed an alternate measure of com-

plexity which avoids any use of probabilistic concepts. Kolmogorov proposed that length of the shortest binary program that generates the sequence be used as a measure of complexity. Motivated by this approach, the *Lempel-Ziv complexity* (Lempel and Ziv, 1976), *finite state complexity* (Ziv and Lempel, 1978a), and *compressibility* (Ziv and Lempel, 1978b) were defined for individual sequences instead of probabilistic sources. Ziv and Lempel (1978a, 1978b) proved coding theorems and their converses for finite-state complexity and compressibility without using the Shannon entropy, and showed that both concepts are closely related to the source's Shannon entropy by proving Eq. (5) for stationary ergodic sources. They also proposed expanding the concept of entropy to nonstationary sources by using the expected compressibility as an analog of entropy. Verdú and Han (1997) proved that satisfying the entropy theorem is equivalent to satisfying the noiseless coding theorem. They also proposed an extension to the definition of entropy under which some nonstationary sources satisfied the coding theorems. Muramatsu and Kanaya's (1999) extension to the definition of entropy, *almost sure sup entropy*, replaces the limit in Eq. (4) of Definition 2 with a limit supremum. Under this extension of entropy, the coding theorem and its converse hold with probability one for all sources with a consistent probability law.

The basic theory underlying the SWML estimator was developed by Wyner and Ziv (1989). Equation (19) was proven with convergence in probability (Wyner and Ziv, 1989), which was later sharpened to convergence with probability one (Ornstein and Weiss, 1993). Wyner (1993) initially proved the convergence of the average match length [Eq. (20)] for finite memory and fixed window size, which was later expanded in Wyner and Wyner (1995). Kontoyiannis and Suhov (1994) proved a similar result for the Doeblin condition, further developed in Kontoyiannis *et al.* (1998). The SWML entropy estimator was proposed in similar forms by Kontoyiannis, 1997; Kontoyiannis *et al.*, 1998; and Wyner *et al.*, 1998.

[1]This class of encoders are called *noiseless* or *faithful* encoders, meaning that the probability of *any* decoding error is identically zero. The discussion in this paper is limited to this class of encoders, including the well known Ziv and Lempel (1997) encoder. In the literature, faithful encoders are distinguished from the broader class of encoders whose probability of decoding error is finite but can be made arbitrarily small.

[2]A Markov model of order k is equivalent to the $(k+1)$-th order approximation model Shannon (1948, 1951) used for his experiments. In particular, the zeroth order Markov model is equivalent to the i.i.d. model or Shannon's first-order model.

[3]Doeblin condition: There exists an integer $k \geq 1$ and a real number $\gamma \in (0,1)$ such that $\Pr\{X_i = x_i | X_{-\infty}^{i-k}\} \geq \gamma$ for all $x_i \in \mathcal{A}$ at all time i with probability one.

[4]The number of parameters of a Markov model and the required sequence length grow with the number of possible transitions, not combinations. An erroneous guideline using the number of combinations appears in McCowan *et al.* (1999). They claimed that their kth-order Markov models had $C(|\mathcal{A}|, k+1) = |\mathcal{A}|!/(k+1)!(|\mathcal{A}|-(k+1))!$ parameters and required the sequence length to exceed this number. The correct number of parameters is actually $|\mathcal{A}|^{(k+1)}$, and thus much longer sequences are required. For example, when $|\mathcal{A}|=20$, a second-order Markov model has 8000 parameters, almost an order of magnitude more than the 1140 predicted by C(20,3). Entropy estimates from incorrect models or insufficient data are uninterpretable.

[5]The bias of an estimator is defined as $\mathbf{E}\{\hat{\theta}\} - \theta$. A negative bias indicates that the expected estimate is lower than the true value.

[6]In contrast to Payne and McVay's (1971) observation, we found some units significantly longer than 6 s. Therefore, we allocated 12.8 s for the time axis of the spectrogram.

Basharin, G. P. (**1959**). "On a statistical estimate for the entropy of a sequence of independent random variables," Theor. Probab. Appl. **4**, 333–336.

Beecher, M. D. (**1989**). "Signalling systems for individual recognition: an information theory approach," Anim. Behav. **38**, 248–261.

Breiman, L. (**1957**). "The individual ergodic theorem of information theory," Ann. Math. Stat. **28**, 809–811 also **31**, 809–810 (**1960**).

Breiman, L. (**1960**). "A correction to 'The individual ergodic theorem of information theory' ", Ann. Math. Stat. **31**, 809–810.

Chomsky, N. (**1956**). "Three models for the description of language," IRE Trans. Inf. Theory **2**, 113–124.

Chomsky, N. (**1988**). *Language and Problems of Knowledge: The Managua Lectures* (MIT Press, Cambridge, MA).

Cover, T. M., and King, R. C. (**1978**). "A convergent gambling estimate of the entropy of English," IEEE Trans. Inf. Theory **24**, 413–420.

Cover, T., and Thomas, J. (**1991**). *Elements of Information Theory* (Wiley, NY).

D'Vincent, C. G., Nilson, R. M., and Hanna, R. E. (**1985**). "Vocalization and coordinated feeding behavior of the humpback whale in southeastern Alaska," Sci. Rep. Whales Res. Inst. **36**, 41–47.

Duda, R. O., Hart, P. E., and Stork, D. G. (**2001**). *Pattern Classification*, 2nd ed. (Wiley, NY).

Fitch, W. T., and Hauser, M. D. (**2004**). "Computational constraints on syntactic processing in a nonhuman primate," Science **303**, 376–380.

Gentner, T. Q., and Hulse, S. H. (**1998**). "Perceptual mechanisms for individual vocal recognition in European starlings, *Sturnus vulgaris*," Anim. Behav. **56**, 579–594.

Good, I. J. (**1953**). "The population frequencies of species and estimation of population parameters," Biometrika **40**, 237–264.

Guinee, L. N., and Payne, K. B. (**1988**). "Rhyme-like repetitions in songs of humpback whales," Ethology **79**, 295–306.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (**2002**). "The faculty of language: What is it, who has it, and how did it evolve?," Science **298**, 1569–1579.

Haykin, S. (**1999**). *Neural Networks: A Comprehensive Foundation*, 2nd ed., (Prentice–Hall, Englewood Cliffs, NJ).

Janik, V. M. (**1999**). "Pitfalls in the categorization of behaviour: A comparison of dolphin whistle classification methods," Anim. Behav. **57**, 133–143.

Janik, V. M., and Slater, P. J. B. (**1997**). "Vocal learning in mammals," in *Advances in the Study of Behavior, Volume 26*, edited by P. J. B. Slater, C. Snowdon, J. Rosenblatt, and M. Milinksi (Academic, NY), pp. 59–99.

Jankowski, C. R., Vo, H.-D. H., and Lippmann, R. P. (**1995**). "A comparison of signal processing front ends for automatic word recognition," IEEE Trans. Speech Audio Process. **3**, 286–293.

Ketten, D. R. (**1997**). "Structure and function in whale ears," Bioacoustics **8**, 103–135.

Khinchin, A. I. (**1953**). "The entropy concept in probability theory," Usp. Mat. Nauk **8**, 3–20 [English translation in Khinchin (1957)].

Khinchin, A. I. (**1957**). *Mathematical Foundations of Information Theory* (Dover, NY).

Kohonen, T. (**2001**). *Self-Organizing Maps*, 3rd ed., (Springer, NY).

Kolmogorov, A. N., (**1965**). "Three approaches to the quantitative definition of information," Probl. Inf. Transm. **1**, 1–7.

Kontoyiannis, I. (**1997**). "The complexity and entropy of literary styles," NSF Technical Report No. 97, Department of Statistics, Stanford University, Palo Alto, CA, June 1996/October 1997.

Kontoyiannis, I., and Suhov, Y. M. (**1994**). "Prefixes and the entropy rate for long-range sources," in *Probability, Statistics and Optimization*, edited by F. P. Kelly (Wiley, NY), pp. 89–98.

Kontoyiannis, I., Algoet, P. H., Suhov, Y. M., and Wyner, A. J. (**1998**). "Nonparametric entropy estimation for stationary processes and random fields, with applications to English text," IEEE Trans. Inf. Theory **44**, 1319–27.

Lempel, A., and Ziv, J. (**1976**). "On the complexity of finite sequences," IEEE Trans. Inf. Theory **22**, 75–81.

Levitin, L. B., and Reingold, Z. (**1994**). "Entropy of natural languages: Theory and experiment," Chaos, Solitons Fractals **4**, 709–743.

Linde, Y., Buzo, A., and Gray, R. M. (**1980**). "An algorithm for vector quantizer design," IEEE Trans. Commun. **28**, 84–95.
MacKay, D. M. (**1972**). "Formal analysis of communicative processes," in *Nonverbal Communication*, edited by R. A. Hinde (Cambridge University Press, Cambridge).
Marton, K., and Shields, P. (**1994**). "Entropy and the consistent estimation of joint distributions," Ann. Prob. 22, 960–977.
Marton, K., and Shields, P. (**1996**). "Entropy and the consistent estimation of joint distributions," Ann. Prob. 24, 541–545.
McCowan, B., Hanser, S. F., and Doyle, L. R. (**1999**). "Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires," Anim. Behav. **57**, 409–419.
McMillan, B. (**1953**). "The basic theorems of information theory," Ann. Math. Stat. **24**, 196–219.
Miller, G. A. (**1954**). "Communication," Annu. Rev. Psychol. **5**, 401–420.
Miller, G. A., and Chomsky, N. (**1963**). "Finitary models of language users," in *Handbook of Mathematical Psychology*, edited by R. D. Luce, R. R. Bush, and E. Galanter (Wiley, NY), pp. 419–492.
Miller, P. J. O., Biassoni, N., Samuels, A., and Tyack, P. L. (**2000**). "Whale songs lengthen in response to sonar," Nature (London) **405**, 903.
Muramatsu, J., and Kanaya, F. (**1999**). "Almost-sure variable-length source coding theorems for general sources," IEEE Trans. Inf. Theory **45**, 337–342.
Noad, M., Cato, D. H., and Bryden, M. M. (**2000**) "Cultural revolution in whale songs," Nature (London) **408**, 537.
Ornstein, D. S., and Weiss, B. (**1993**). "Entropy and data compression schemes," IEEE Trans. Inf. Theory **39**, 78–83.
Patil, G. P., and Taillie, C. (**1982**). "Diversity as a concept and its measurement," J. Am. Stat. Assoc. **77**, 548–567 (with discussions).
Payne, R. (**1995**). *Among Whales* (Scribner's Sons, NY).
Payne, R. S., and McVay, S. (**1971**). "Songs of humpback whales," Science **173**, 587–597.
Payne, K., Tyack, P., and Payne, R. (**1983**). "Progressive changes in the songs of humpback whales (*Megaptera novaeangliae*): A detailed analysis of two seasons in Hawaii," in *Communication and Behavior of Whales*, edited by R. Payne, AAAS Selected Symposium 76 (Westview, Boulder, CO), pp. 9–58.
Peet, R. K. (**1974**). "The measurement of species diversity," Annu. Rev. Ecol. Syst. **5**, 285–307.
Pitton, J. W., Wang, K., and Juang, B.-H. (**1996**) "Time-frequency analysis and auditory modeling for automatic recognition of speech," Proc. IEEE **84**, 1199–1215.
Savage-Rumbaugh, S., McDonald, K., Sevcik, R. A., Hopkins, W. D., and Rubert, E. (**1986**), "Spontaneous symbol acquisition and communicative use by pygmy chimpanzees (*Pan paniscus*)," J. Exp. Psychol. **115**, 211–235.
Shannon, C. E. (**1948**). "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423.
Shannon, C. E. (**1951**). "Prediction and entropy of printed English," Bell Syst. Tech. J. **30**, 50–64.
Shields, P. C. (**1996**). *Ergodic Theory of Discrete Sample Paths* (American Mathematical Society, Providence, RI).
Slater, P. J. B. (**1973**). "Describing sequences of behavior," in *Perspectives in Ethology*, edited by P. P. G. Bateson, and P. H. Klopfer, (Plenum Press, New York), Vol. 1, pp. 131–153.
Tyack, P. L. (**1981**). "Interactions between singing Hawaiian humpback whales and conspecifics nearby," Behav. Ecol. Sociobiol. **8**, 105–116.
Tyack, P. L. (**1998**). "Acoustic communication under the sea," in *Animal Acoustic Communication: Sound Analysis and Research Methods*, edited by S. L. Hopp, M. J. Owren, and C. S. Evans (Springer, Berlin), pp. 163–220.
Tyack, P. L., and Sayigh, L. S. (**1997**). "Vocal learning in cetaceans," in *Social Influences on Vocal Development*, edited by C. Snowdon, and M. Hausberger (Cambridge University Press, Cambridge), pp. 208–233.
Verdú, S., and Han, T. S. (**1997**). "The role of the asymptotic equipartition property in noiseless source coding," IEEE Trans. Inf. Theory **43**, 847–857.
Walker, A., Fisher, R. B., and Mitsakakis, N. (1996). "Singing maps: Classification of whalesong units using a self-organizing feature mapping algorithm," Research Paper No. 833, Department of AI, University of Edinburgh, UK.
Winn, H. E., and Winn, L. K. (**1978**). "The song of the humpback whale *Megaptera novaeangliae* in the West Indies," Mar. Biol. (Berlin) **47**, 97–114.
Wyner, A. J. (1993). "String matching theorems and applications to data compression and statistics," Ph.D. dissertation, Department of Statistics, Stanford University, Palo Alto, CA.
Wyner, A. D., and Ziv, J. (**1989**). "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression," IEEE Trans. Inf. Theory **35**, 1250–1258.
Wyner, A. D., and Wyner, A. J. (**1994**). "The sliding window Lempel–Ziv algorithm is asymptotically optimal," Proc. IEEE **82**, 872–877.
Wyner, A. D., and Wyner, A. J. (**1995**). "Improved redundancy of a version of the Lempel–Ziv algorithm," IEEE Trans. Inf. Theory **41**, 723–731.
Wyner, A. D., Ziv, J., and Wyner, A. J. (**1998**). "On the role of pattern matching in information theory," IEEE Trans. Inf. Theory **44**, 2045–2056.
Ziv, J., and Lempel, A. (**1977**). "A universal algorithm for sequential data compression," IEEE Trans. Inf. Theory **23**, 337–343.
Ziv, J., and Lempel, A. (**1978a**). "Coding theorems for individual sequences," IEEE Trans. Inf. Theory **24**, 405–412.
Ziv, J., and Lempel, A. (**1978b**). "Compression of individual sequences via variable-rate coding," IEEE Trans. Inf. Theory **24**, 530–536.

Acoustic features of objects matched by an echolocating bottlenose dolphin

Caroline M. DeLong[a] and Whitlow W. L. Au
Hawaii Institute of Marine Biology, P. O. Box 1106, Kailua, Hawaii 96734

David W. Lemonds
Lockheed Martin Orincon, 970 North Kalaheo Avenue, Suite C-215, Kailua, Hawaii 96734

Heidi E. Harley
The Living Seas, Epcot®, Walt Disney World® Resort, New College of Florida, 5700 North Tamiami Trail, Sarasota, Florida 34243

Herbert L. Roitblat
DolphinSearch, Inc., 5855 Olivas Park Drive, Ventura, California 93003

(Received 26 April 2005; revised 27 October 2005; accepted 3 December 2005)

The focus of this study was to investigate how dolphins use acoustic features in returning echolocation signals to discriminate among objects. An echolocating dolphin performed a match-to-sample task with objects that varied in size, shape, material, and texture. After the task was completed, the features of the object echoes were measured (e.g., target strength, peak frequency). The dolphin's error patterns were examined in conjunction with the between-object variation in acoustic features to identify the acoustic features that the dolphin used to discriminate among the objects. The present study explored two hypotheses regarding the way dolphins use acoustic information in echoes: (1) use of a single feature, or (2) use of a linear combination of multiple features. The results suggested that dolphins do not use a single feature across all object sets or a linear combination of six echo features. Five features appeared to be important to the dolphin on four or more sets: the echo spectrum shape, the pattern of changes in target strength and number of highlights as a function of object orientation, and peak and center frequency. These data suggest that dolphins use multiple features and integrate information across echoes from a range of object orientations. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2161434]

PACS number(s): 43.80.Ka, 43.80.Lb [FD] Pages: 1867–1879

I. INTRODUCTION

Dolphins echolocate by emitting short, high-intensity clicks and processing the echoes reflected from objects. The ability of dolphins to detect and discriminate among objects via echolocation is well documented (for a recent review, see Au, 2000). Although these capabilities have been characterized, it is still not clear which acoustic features of object echoes are information-bearing parameters for dolphins, i.e., which features convey object properties such as size, shape, and material to an echolocating dolphin.

In a number of previous studies, experimenters have presented echolocating dolphins with discrimination tasks and then examined the echoes from the objects used in those tasks with simulated dolphin sonar signals (e.g., Au and Martin, 1988; Au and Pawloski, 1992; Au and Turl, 1991; Hammer and Au, 1980; Nachtigall *et al.*, 1980). An inspection of the object echoes allowed the experimenters to speculate about the echo features the dolphins may have used to discriminate between the objects. For example, Hammer and Au (1980) investigated a dolphin's ability to discriminate between hollow cylinders of the same diameter that varied in material composition (aluminum, bronze, glass, and steel). The dolphin performed well, and after the task was completed the cylinder echoes were obtained using simulated dolphin signals. A visual inspection of the cylinder echoes showed that the targets each had different arrival times for the secondary echo "highlight" (local maximum in echo amplitude). The researchers suggested that the predominant cue used by the dolphin in discriminating among the cylinders was probably time separation pitch generated by the first and second highlights. Humans, when presented with a pair of correlated sound pulses, perceive a pitch equal to $1/T$, where T is the time separation between pulses (Small and McClellan, 1963; McClellan and Small, 1965).

Such studies have been useful in determining a number of echo features that are *available* to dolphins asked to discriminate among objects that vary in size, shape, material, or structure (e.g., highlight structure of the echo waveform, time separation pitch, target strength, and frequency shifts in the peaks and nulls of the echo spectrum). However, it is not clear that dolphins actually attend to and utilize these echo features to give them information about object characteristics.

[a]Present address: Brown University, Dept. of Neuroscience, Box 1953, Providence, RI 02912. Electronic mail: Caroline_Delong@Brown.edu

One way to determine which echo features dolphins actually use during a discrimination task is to take a quantitative approach to analyzing the object echoes. Instead of a visual inspection of the object features, the between-object differences in each of these features can be measured. In addition, the dolphin's error patterns during the discrimination task (e.g., how often each object is confused with each other object) can be analyzed in conjunction with the between-object variation in acoustic features to identify the acoustic features that the dolphin may have used. Two objects are confused to the degree that they share similar features. The combination of error patterns and acoustic similarity patterns thus indicate the specific features that underlie the confusion. For example, imagine that object 1 and object 2 had echoes that were very close in amplitude, but objects 1 and 3 had echoes that were significantly different in amplitude. If the dolphin confused objects 1 and 2 but did not confuse objects 1 and 3, then it can be inferred that the dolphin may have used amplitude to make its decision. In contrast, if a dolphin did not confuse objects 1 and 2, but confused objects 1 and 3, then it suggests that the dolphin may have used a different acoustic feature to make its decision. This method of comparing the dolphin's error patterns with the acoustic similarities between the echoes was employed in the present study to investigate how the dolphin used the acoustic features of the echoes.

The approach of this study was to give a dolphin an echolocation matching task in which he was presented with a sample object and then had to choose the identical object from among three alternatives. The dolphin was presented with a variety of objects that differed along one or more dimensions (e.g., size, shape, material, texture). After the matching task was completed, the object echoes were recorded in a test tank by projecting a dolphin click at the objects and the acoustic features of the echoes were measured. This study explored two hypotheses regarding the way dolphins use acoustic information in echoes: use of a single feature (hypothesis no. 1), or use of a linear combination of multiple features (hypothesis no. 2). Since there is no one-to-one correspondence between object characteristics (e.g., size, shape), and echo features (e.g., target strength, number of highlights), we predicted that the dolphin would use a combination of multiple acoustic features instead of a single acoustic feature.

II. BOTTLENOSE DOLPHIN ECHOIC MATCHING EXPERIMENT

A. Animal subject

The subject was an adult male Atlantic bottlenose dolphin (*Tursiops truncatus*) housed at Disney's Epcot's Living Seas in Orlando, Florida. At the beginning of the study the dolphin Toby was approximately 20 years old. Toby was an experienced research subject (Bauer and Johnson, 1994; Xitco, 1996) and had extensive experience with echoic matching (Xitco and Roitblat, 1996) and cross-modal matching (Harley *et al.*, 2003). Sessions were conducted in the main tank of the Living Seas (circular salt-water aquarium about 67 m in diameter and 9 m deep, with a volume of 22 million liters). During sessions, Toby usually received one-quarter of his daily allotment of approximately 9.5 kg of fish (herring [*Culpea harengus*], mackerel [*Scomber japonicus*], and capelin [*Mallotus villosus*]).

B. Materials and procedure

The stimuli were 27 assorted hardware-store objects that were unfamiliar to the dolphin (Fig. 1). The objects varied in size from 4.8 cm height $\times$ 4.8 cm width $\times$ 1.3 cm depth (smallest object: Small Stone Square) to 38.1 cm $\times$ 20.7 cm $\times$ 9.8 cm (largest object: Large Strainer). The objects were organized into nine object sets and each set contained three objects. Objects within each set were selected to vary along one nominal dimension: size, shape, material, or texture. Two object sets were selected for size differences (Strainers and Stone Squares), two object sets were selected for shape differences (Foam Cones and Stone Shapes), two object sets were selected for material differences (Rods and Figure 8's), and three object sets were selected for texture differences (Wooden Plaques, Green Foam, and Sockets). However, the objects were natural stimuli so they did not vary along only one dimension. For example, the Stone Shapes varied in both shape and size.

Samples were presented in the underwater sample apparatus, a 0.7 m square PVC-framed box tightly covered in 10 mm black polyethylene that was acoustically transparent but visually opaque. Comparison stimuli were presented in the choice apparatus, a larger (front: 2.76 m by 0.90 m, sides: 0.98 m by 0.90 m) but similarly fashioned rectangular structure, the top of which was positioned one meter below the water's surface (also covered in black polyethylene). The three comparison stimuli were centered within three square sections in the rectangular structure.

The basic procedure involved a three-alternative identity match-to-sample task in which the dolphin was presented with a sample stimulus, the sample was removed, and then the dolphin was required to select the identical stimulus from among three comparison stimuli. The objects were only accessible to the dolphin using echolocation. At the beginning of each trial, the dolphin investigated the sample as long as he wished at the underwater sample apparatus, after which he swam to the choice array located several meters behind him. When inspecting both the sample and the choices, the dolphin's head was unrestrained and he was free to ensonify the objects from several orientations (although he was not allowed to swim behind the sample or choice apparatus). After the dolphin positioned himself in front of his object of choice for about 3 s an assistant naïve to the sample's identity reported the dolphin's choice to the trainer who blew a whistle and reinforced the dolphin with two small fish for a correct choice, or tapped on a metal platform to recall him for an incorrect choice. Intertrial intervals averaged approximately 60 s (minimum 30 s).

Object sets were presented in the following order: Strainers, Wood Plaques, Foam Cones, Figure 8's, Stone Squares, Stone Shapes, Sockets, Green Foam, and Rods. Two 18-trial sessions were presented for each object set. Within a session, each stimulus was presented as the sample object an

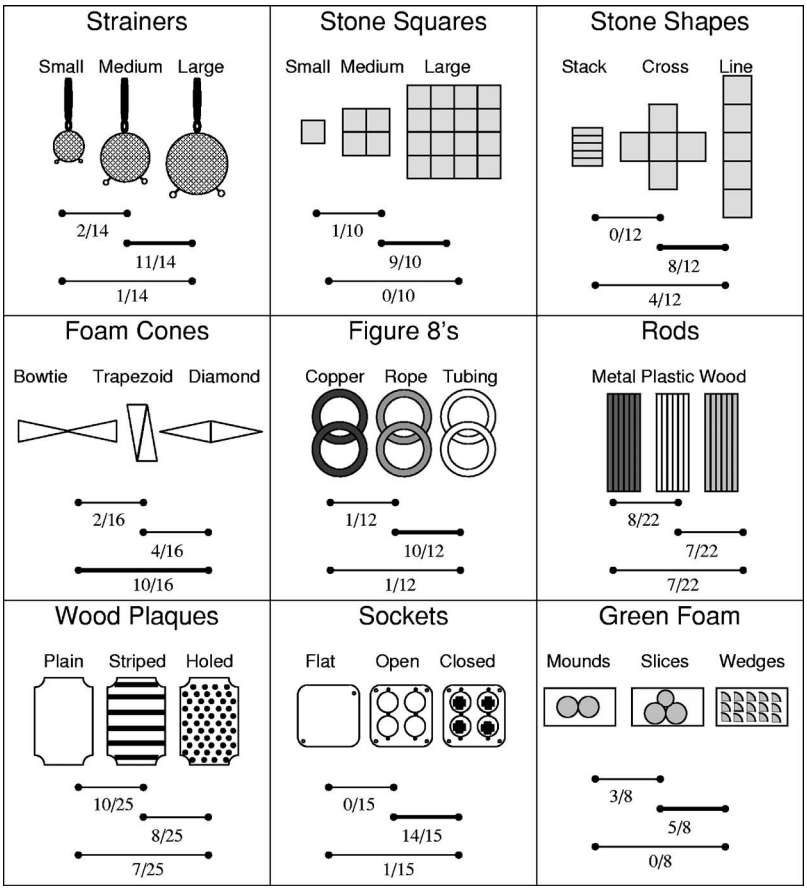


FIG. 1. Object sets. The dolphin's errors are shown for each object pair by lines connecting those objects. The number of errors the dolphin made for each pair is given under the line. The bold line indicates the "predominant" error (see text for details).

equal number of times. Each object was also presented equally often in each choice position. The order of the trials was randomized. The dolphin's ability to discriminate among the objects was measured using both percent correct matches and the unbiased sensitivity parameter, *d*' (Green and Swets, 1988).[1] The use of *d*' is preferable to percent correct matches because discrimination is inferred from nonrandom responding given the comparison choices, but does not require the subject to select the matching choice alternative. In other words, the dolphin could still show evidence of being able to discriminate the objects, even if it employs a perverse response rule (e.g., choosing object 2 every time object 1 is presented).

C. Results and discussion

Overall choice accuracy for all nine sets is as follows: Strainers (61%), Stone Squares (72%), Stone Shapes (67%), Foam Cones (56%), Figure 8's (67%), Rods (39%), Wood Plaques (31%), Sockets (58%), and Green Foam (78%). Chance choice accuracy is 33%, because the dolphin could choose from among three alternatives. The alpha level used to determine significance throughout this article was 0.05. A binomial test rather than parametric statistics was used because of the single subject design. The dolphin's choice accuracy was significantly above chance for all object sets except for the Rods and Wood Plaques (summed binomial test).

Figure 1 displays the dolphin's errors for each object set (under the lines connecting each pair of objects). In a match-to-sample task, errors occur when the dolphin chooses an object that does not match the sample object. For example, when the sample is "Large" and the dolphin chooses "Medium" the dolphin has made a Large-Medium error (Large-Medium and Medium-Large errors are collapsed into one group and shown under the line connecting those objects in Fig. 1). Chi-square tests were used to determine whether the dolphin's errors were distributed uniformly among the three possible object confusions. In the seven object sets in which the dolphin's overall choice accuracy was above chance, it confused one of the three object pairs more frequently than the other two pairs (Stone Squares $[\chi^2(2)=14.6, N=10, p<0.001]$, Strainers $[\chi^2(2)=13.0, N=14, p<0.01]$, Stone Shapes $[\chi^2(2)=8.0, N=12, p<0.05]$, Foam Cones $[\chi^2(2)=6.5, N=16, p<0.05]$, Figure 8's $[\chi^2(2)=13.5, N=12, p<0.01]$, Sockets $[\chi^2(2)=24.4, N=15, p<0.001]$, Green Foam $[\chi^2(2)=6.8, N=8, p<0.05]$).

The pair of objects the dolphin confused most often was called the predominant error (defined as the error type that accounts for more than 50% of the errors, and shown as bold lines in Fig. 1). For example, in the Stone Squares set, the dolphin's predominant error was to confuse the Medium and Large objects. In the other two object sets, Rods and Wood Plaques (overall choice accuracy below chance), the dolphin's errors were distributed uniformly among the three possible object confusions (Rods $[\chi^2(2)=0.1, N=22, p>0.05]$, Wood Plaques $[\chi^2(2)=0.6, N=25, p>0.05]$). For these sets, there was no predominant error because the dolphin confused all three pairs of objects.

III. ACOUSTIC MEASUREMENTS OF OBJECTS

A. Materials and procedure

After the completion of the cross-modal matching experiment with the dolphin, echoes from all the objects were obtained using a representative click recorded from a different male bottlenose dolphin. This dolphin click, which has been used in numerous studies (see Au, 1993), was 70 μs long with a peak frequency of about 120 kHz and a 60-kHz bandwidth. Dolphin clicks can vary in frequency and intensity, but this click was considered to be a typical, average click that was likely to be very similar to the clicks that were used by the dolphin Toby during this study (Au, 1993). The object echoes were obtained using this recorded dolphin click because recording real-time echoes from the dolphin's own clicks during the matching task was not feasible given the time constraints of this study (see general discussion).

The measurements took place in a seawater tank (with the exception of the extremely buoyant Foam Cones which were measured in an open ocean pen but otherwise in a similar fashion; see Fig. 1 in Benoit-Bird *et al.*, 2003). The cylindrical tank was 1.82 m in height by 2.41 m in diameter and contained 8.3 m^3 of seawater. The transmitting transducer and the receiving hydrophone were mounted on the same transducer assembly (a 28.55 cm $\times$ 20.30 cm aluminum plate). The transmitting transducer was located 2.54 cm above the receiver. The transducer and receiver were custom built with piezoelectric ceramic circular disks that were 6.35 cm in diameter and 6.35 mm thick (Material Systems).

The dolphin click signal was generated by a Quatech WSB-10 function generator board housed in a PC and amplified (Hafler P3000 Transnova). The received echoes were gated, amplified using a custom-built amplifier, and filtered before being digitized at 1 MHz using a Rapid Systems R1200 A/D Converter. An oscilloscope was used to view the signals during echo collection (Tektronix TDS 210).

The transducer assembly and the targets were placed 1 meter below the water's surface. Individual targets were hung with monofilament line from a T-bar suspended over the surface of the water. The T-bar was linked to a calibrated rotor (ILC Data Device Corp. API 30602) that could rotate the targets 360 degrees in 1.3 degree increments. Some of the objects required lead weights (343 or 571 g) attached with monofilament line to make them negatively buoyant. These weights were hung 62.23 cm below the targets to minimize their contribution to the echoes. These weights were not hung from the targets during the dolphin echoic matching task. Weights were not needed during the echoic matching task because the targets were held in place in the target boxes using thin polyethylene strings attached to the top and bottom of the targets (not possible in the measurement tank).

In order to simulate the experience the dolphin could get from swimming by the target and ensonifying it from different angles, echoes were measured from multiple angles parallel to the horizontal axis of the target. Multiple aspect angles were measured because the dolphin could ensonify the object from different angles relative to the object during the matching task and because there is evidence that dolphins attend to the pattern of changes in acoustic features as an object is scanned across a range of target orientations (e.g., Nachtigall *et al.*, 1980).

It was estimated that the dolphin was exposed to a range of angles spanning +/− 15 degrees (30 degree span) during the matching task (videotapes of some sessions suggest that the dolphin did not consistently investigate a wider range of angles). Simulating this range, each measurement of an object produced a 23-echo train in which one echo was captured for each angle (1.3 degrees apart) between −15 degrees and +15 degrees (the orientation of the target that faced the front of the choice apparatus in the matching task was called the 0 degree angle). Ten echo train measurements were collected for each object (23 echoes per train $\times$ 10 trains per object $\times$ 27 objects = 6210 total echoes). All ten measurements for a single object were completed consecutively on the same day (all ten measurements were made for one object before measuring another object). The objects were measured in random order.

1. Acoustic feature extraction

Six acoustic features were extracted from each individual echo: (1) target strength, (2) number of highlights, (3) duration, (4) peak frequency, (5) center frequency, and (6) rms bandwidth (Figs. 2(a) and 2(b)). These features characterize the echo in the time and frequency domains and are standard features that have been examined in other studies (for a review see Au, 1993).

Target strength is defined as the ratio in dB of the echo intensity measured 1 m from the target to the intensity of the incident signal at the location of the target. Highlights, defined as local amplitude maxima in the time domain, were localized and counted by finding the envelope of the signal, smoothing the envelope (by averaging 15 points centered around the sample), thresholding the smoothed envelope, taking the derivative of the thresholded envelope, representing the derivative as a trinary signal (+1, 0, −1), taking the derivative of the trinary signal, and extracting zero crossings that indicate slope changes associated with highlight onset and peak events. The number of highlights was simply defined as the number of peaks detected.

Echo duration was calculated by finding the envelope of the signal, thresholding the envelope at 1 standard deviation, then finding the time between the first and last suprathreshold sample. The peak frequency is the frequency of the signal at which the spectrum has its maximum value. The center frequency is defined as the frequency that divides the power spectrum into two equal energy parts. The root mean square (rms) bandwidth indicates the frequency range around the center frequency in which the majority of spectral energy lies.

Since there is evidence that dolphins attend to the pattern of changes in acoustic features as an object is scanned across a range of orientations (Nachtigall *et al.*, 1980), additional acoustic features were extracted from echo trains instead of individual echoes (one train is 23 echoes collected from −15 to +15 degrees; Figs. 2(c) and 2(d)). Four features were calculated: relative target strength, relative number of highlights, target strength "bumpiness," and highlight

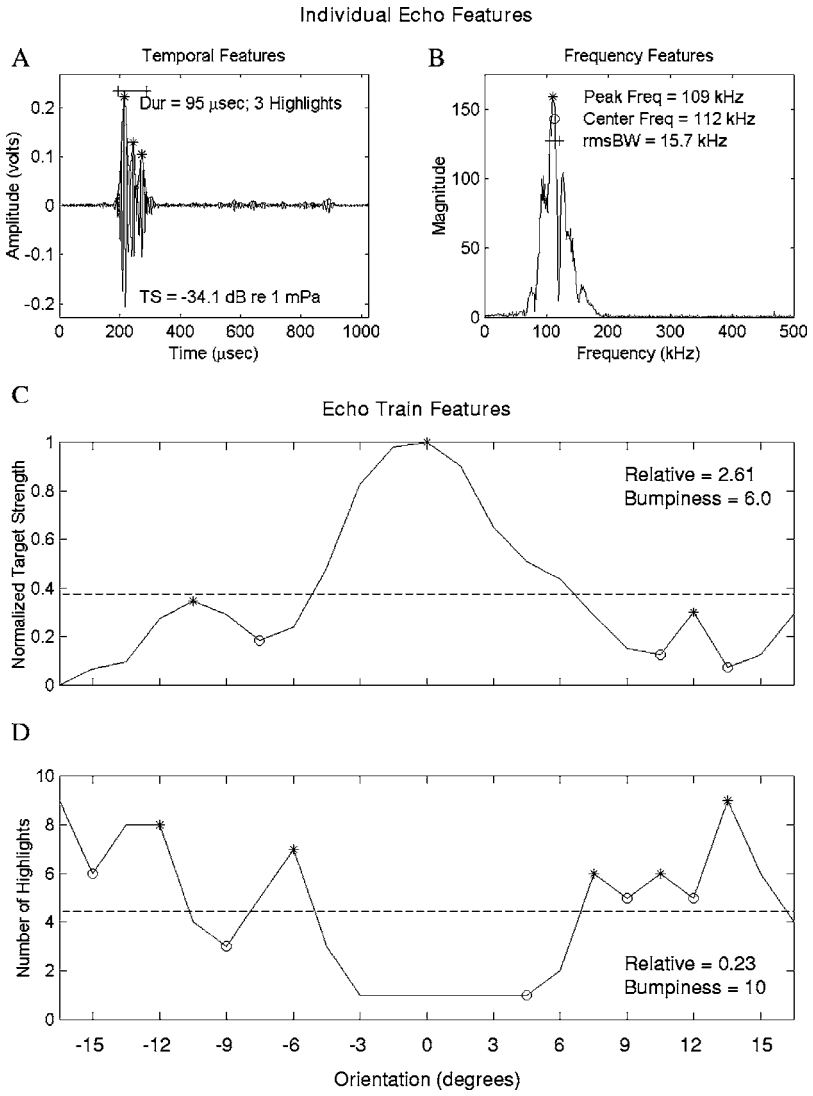


FIG. 2. Features that were extracted from the echoes. Parts A and B show features extracted from individual echoes. This example echo is from the Cross. Part A shows the echo in the time domain and the two temporal features that were extracted: duration (line above echo) and number of highlights (marked with asterisks). Target strength is also shown on the bottom of the graph. Part B shows the echo in the frequency domain and the three frequency features that were extracted: peak frequency, center frequency, and rms bandwidth. Parts C and D show features extracted from echo trains. These example data are from the Flat Socket. Part C shows the normalized target strength plotted as a function of orientation. Part D shows number of highlights plotted as a function of orientation. The dotted line shows the average target strength (or number of highlights) from −15° to +15°. The relative target strength (or number of highlights) score is calculated by dividing the value at the 0° orientation by the value averaged from −15° to +15°. The stars and open circles show changes in line slope. The total number of changes in slope across −15° to +15° yields the "bumpiness" score.

"bumpiness." These features were extracted from each of the ten echo trains for each object. The relative target strength (or number of highlights) score is calculated by dividing the value at the 0 degree orientation by the average value for all orientations. These two features provide a measure of target strength or highlight variability relative to the 0 degree aspect. Target strength (or highlight) bumpiness was defined as the number of slope direction changes across all orientations (see Figs. 2(c) and 2(d)). These echo train features have not been systematically examined in other studies.

B. Results and discussion

1. Variation in single acoustic features of the echoes

a. Changes in acoustic features of the echoes within objects as a function of orientation. Figure 3 shows changes in the echo structure in the time domain as a function of orientation for one of the object sets—the Stone Squares. The way the echo changed as a function of orientation depended on object type. Since echo highlights are caused by reflections from different parts of the target, more complexly shaped targets produce echoes with more complex structures. Whether the waveforms were short and simple or long and complex depended on the number of reflective surfaces presented by the object at each particular angle.

To examine whether acoustic features changed significantly as a function of orientation, separate multivariate analyses of variance (MANOVA) were performed for each object with one independent variable (orientation) and six dependent variables (target strength, number of highlights, peak frequency, center frequency, bandwidth, and duration). A significant multivariate effect of orientation was found for all 27 objects (for all objects, $df=132, 1242$, $p<0.001$). All six acoustic features varied significantly as a function of orientation for all 27 objects (for all objects, $df=22, 230$, $p<0.001$). Each of the acoustic features changed in different ways as a function of orientation, and the patterns of change also varied between objects. An examination of all the object sets revealed that there were many different ways the acoustic features could change as a function of orientation. For example, in the Stone Squares set, target strength increased as the objects rotated toward 0 degrees, whereas the number of highlights and duration either decreased at 0 degrees or remained constant across orientations (see Fig. 4).

b. Acoustic differences between the objects. To examine between-object differences in the acoustic features, separate

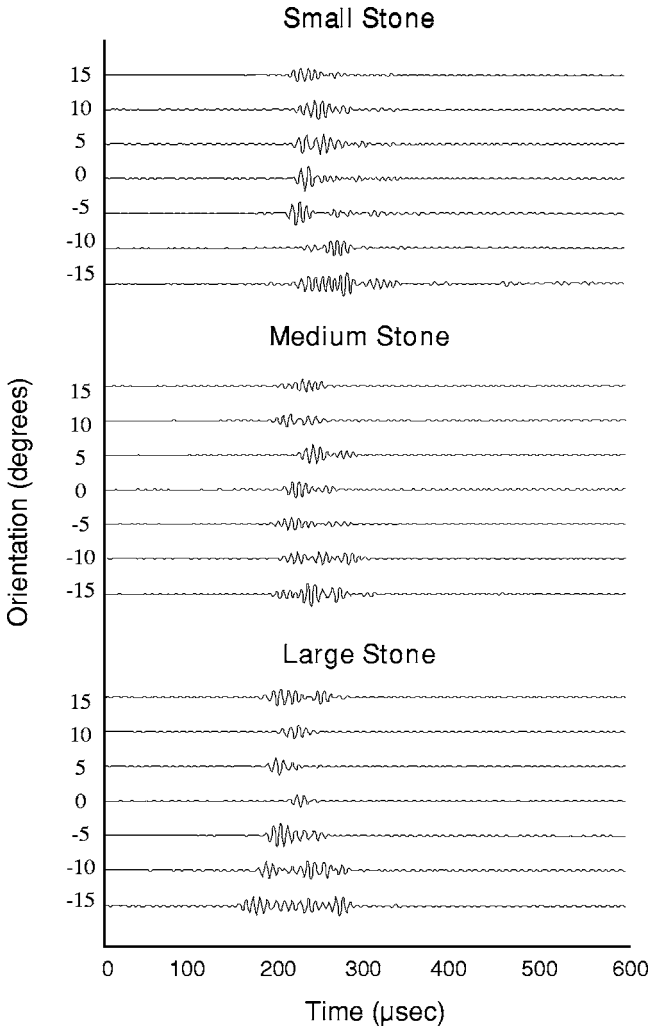


FIG. 3. Echo waveforms for the Stone Squares set showing changes in the echoes of each object as a function of orientation. The orientation facing the front of the choice apparatus during the echoic matching task is called 0°. Amplitude was normalized so amplitude differences between the three objects cannot be seen.

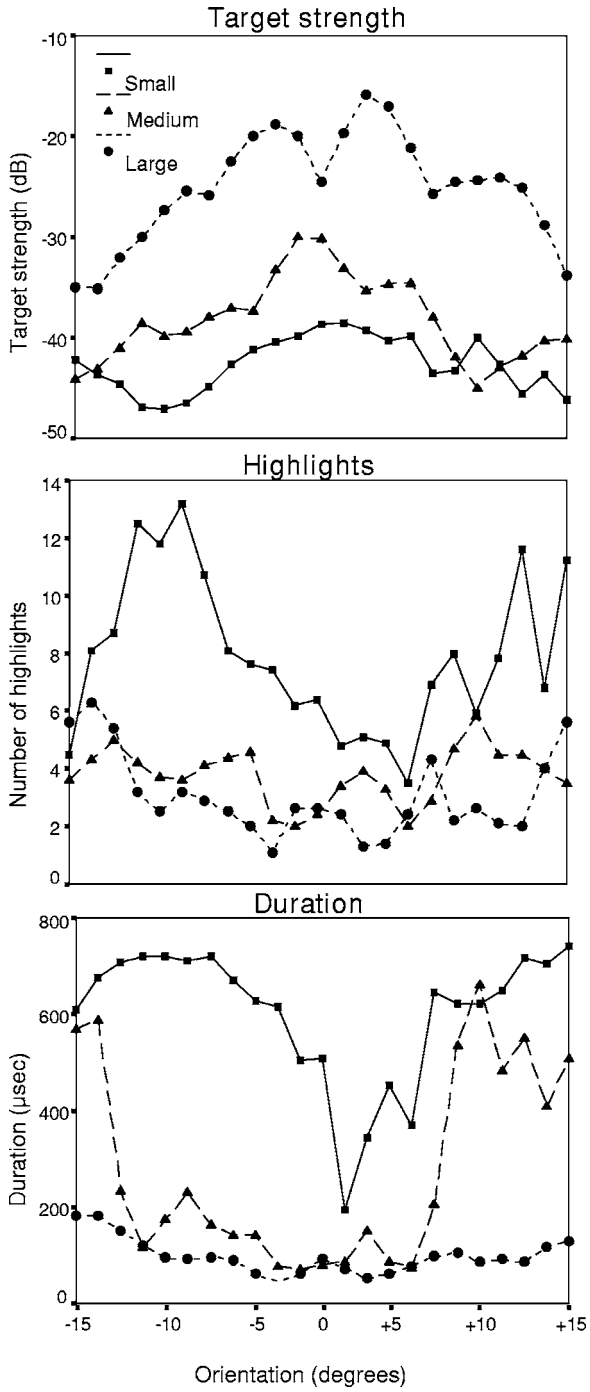


FIG. 4. Changes in three echo features (target strength, number of highlights, duration) as a function of object orientation for the Stone Squares set. The orientation facing the front of the choice apparatus during the echoic matching task is called 0°.

multivariate analyses of variance (MANOVA) were conducted for each of the nine object sets. Between-object differences were examined within object sets (instead of examining differences between all possible pairings of the 27 total objects) because the dolphin was asked to match objects within sets but not between sets. Two separate groups of MANOVAs were conducted for (1) the six acoustic features of individual echoes and (2) the four acoustic features of echo trains. The features of the individual echoes were not analyzed together with the features of the echo trains because the number of samples was different for individual echoes versus echo trains (230 samples versus 10 samples). For the individual echo analysis, acoustic features were extracted from 230 individual echoes per object, one echo from each angle in the range of −15 to +15 degrees (23 echoes) $\times 10$ trains. For the echo train analysis, acoustic features were extracted from 10 echo trains per object (i.e., 10 samples per object).

In the echo MANOVAs, the independent variables were object and orientation and the dependent variables were the six acoustic features of individual echoes. There were significant multivariate effects of object, orientation, and object $\times$ orientation for all nine object sets (for all tests, $p < 0.001$). The acoustic features of the objects differed among the different objects, and the differences among objects changed as the orientation of those objects changed. For seven of the nine object sets (Stone Squares, Strainers, Stone Shapes, Foam Cones, Rods, Sockets, Green Foam), there were significant differences among objects for all six acoustic features (for all features in all sets, $p < 0.001$). For the Figure 8's set, the objects were not significantly different in

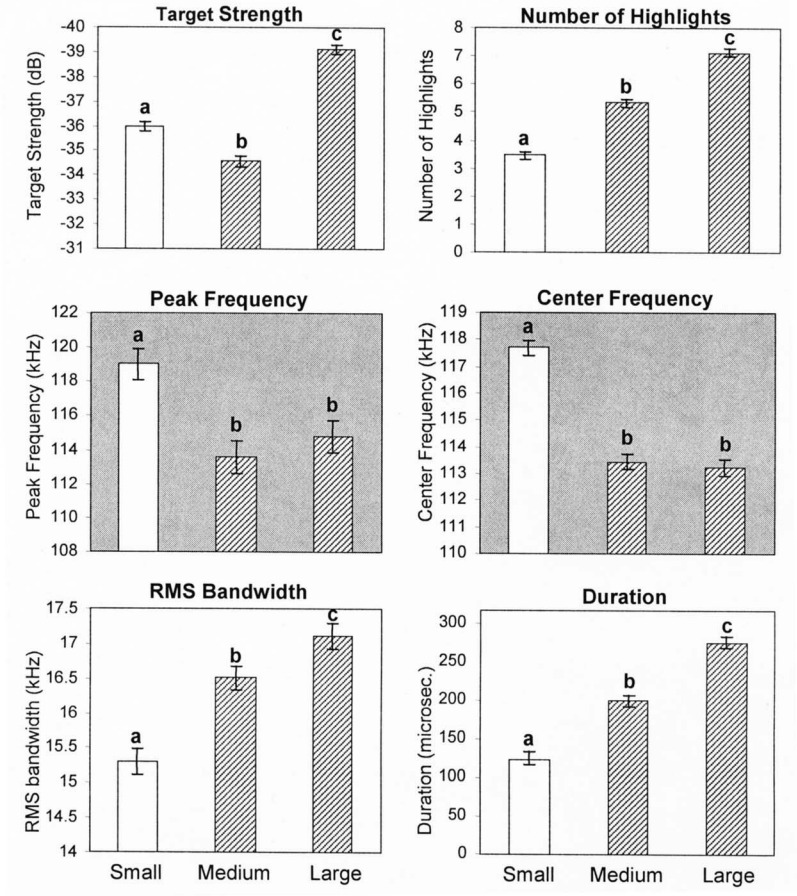


FIG. 5. Features of individual echoes for the Strainers set. The bars representing the objects that the dolphin confused most often (predominant error) are cross-hatched (Medium and Large). Objects sharing the same letter were not statistically significantly different from each other ($p<0.05$). Panels with features that can account for the dolphin's errors are shaded gray.

bandwidth, but the objects were significantly different across all other features ($p<0.001$). For the Wood Plaques set, the objects were not significantly different in peak frequency, but the objects were significantly different across all other features ($p<0.001$).

To examine which pairs of objects within each set showed significant differences in these acoustic features, post-hoc object comparisons were conducted. Figure 5 displays the results of the post-hoc analyses for one of the nine object sets: the Strainers. The height of the vertical bars represents the values for each feature averaged across all object orientations. In Figs. 5 and 6, statistically significant differences between means are indicated by different lower case letters. For example, in the Strainers set, the Small Strainer, Medium Strainer, and Large Strainer are all significantly different in target strength, number of highlights, bandwidth, and duration. The Medium Strainer and the Large Strainer are not significantly different in peak frequency and center frequency, but both are different from the Small Strainer.

To examine which acoustic features the dolphin may have used to discriminate among the objects, the errors made by the dolphin were considered in conjunction with the results of the post-hoc object comparisons. In Fig. 5, the bars representing the objects that the dolphin confused most often (predominant errors) are cross-hatched. If two objects in a set had similar values for a certain acoustic feature (i.e., they were *not* significantly different on that feature), *and* the dolphin made errors between those two objects, then it was inferred that those acoustic features may have been part of the dolphin's decision making process and used by the dolphin to construct representations of the objects. In Fig. 5, graphs with acoustic features inferred to be part of the dolphin's decision making process according to the above logic are shaded in gray. For example, in the Strainers set, the dolphin confused the Medium and Large Strainer, which are similar in peak and center frequency. Thus, peak and center frequency may have been important acoustic features for the dolphin for that object set.

A second group of MANOVAs were conducted to investigate between-object differences in acoustic features calculated from echo trains. A separate MANOVA was conducted for each object set with one independent variable (object) and four dependent variables (relative target strength, relative number of highlights, target strength bumpiness, highlight bumpiness). A significant multivariate object effect was found for all nine object sets (for all sets $df=8, 50$, $p<0.001$). For four of the nine object sets (Strainers, Figure 8's, Rods, Green Foam), there were significant differences among objects for all four acoustic features ($df=2, 27$, $p<0.01$). For three object sets (Stone Squares, Foam Cones, Sockets), the objects were not significantly different in highlight bumpiness (the three other features were significantly different, $p<0.01$). For the final two sets (Stone Shapes, Wood Plaques), the objects were not significantly different in target strength bumpiness (the three other features were significantly different, $p<0.001$). To examine which pairs of objects within each set showed significant differences in these acoustic features, again post-hoc object comparisons

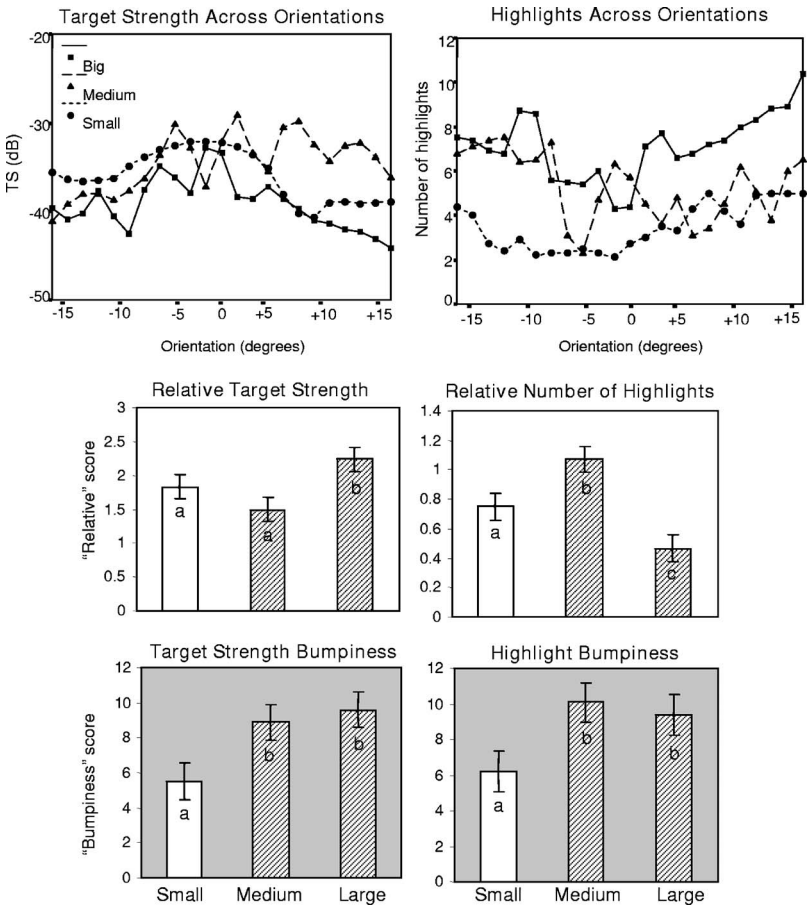


FIG. 6. The top two graphs show the change in target strength and number of highlights as a function of orientation for the Strainers set. The bottom four graphs show four features extracted from the echo trains. The bars representing the objects that the dolphin confused most often (predominant error) are cross-hatched (Medium and Large). Objects sharing the same letter were not statistically significantly different from each other ($p < 0.05$). Panels with features that can account for the dolphin's errors are shaded gray.

were conducted and considered in conjunction with the dolphin's errors. Figure 6 shows the results for the Strainers set.

Table I summarizes the features that may explain the dolphin's errors. Note that different acoustic feature(s) were implicated to be important to the dolphin for each object set. There was not a single acoustic feature or two that seemed to be used by the dolphin for every object set. These results suggest that the dolphin did not use a single acoustic feature to discriminate among all types of objects (hypothesis no. 1). Four features appeared to be important to the dolphin on four or more object sets: target strength bumpiness, highlight bumpiness, peak frequency, and center frequency. In eight of the nine object sets, the dolphin's error patterns implied the use of multiple acoustic features.

c. Comparison of variance within objects to variance between objects. The variability in acoustic features within objects (as a function of orientation) was compared to the variability in acoustic features between objects. The variance

TABLE I. Acoustic features that may explain the dolphin's errors for each of the object sets.

	Object set									
Acoustic features	Stone squares	Strainers	Stone shapes	Foam cones	Figure 8's	Rods	Wood	Sockets	Green foam	Total
Individual Echo										
Target strength			X	X						2
Highlights (#)				X						1
Peak frequency		X		X	X	X	X		X	6
Center frequency		X		X			X	X	X	5
RMS bandwidth			X	X						2
Duration				X						1
Echo Train										
Relative TS			X						X	2
Relative HL					X	X		X		3
TS Bumpiness		X		X		X	X		X	5
HL Bumpiness		X			X	X	X			4
Total	0	4	3	7	3	4	4	2	4	
Dolphin's Accuracy	72%	61%	67%	56%	67%	39%	31%	58%	78%	

was computed for each three-object set ("between-object variance") and then for each object within each set separately ("within-object variance"), with each echo as a single observation. A ratio of within:between object variance was calculated by dividing within-object variance by between-object variance. If the value of this ratio was greater than 1.0, then the within-object variance was greater than the between-object variance.

For one of the object sets in which the dolphin performed significantly above chance (Stone Shapes: 67% correct), the within-object variance was greater than or equal to the between-object variance (ratios ranged from 1.30 to 0.90) for four features for Cross (target strength, number of highlights, center frequency, bandwidth), five features for Line (target strength, peak and center frequency, bandwidth, duration), and three features for Stack (number of highlights, peak frequency, duration). These results show the dolphin succeeded at the matching task even when the acoustic variance within objects (as a function of orientation) was greater than or equal to the variance between objects.

2. Spectral correlations

To test the hypothesis that the shape of the echo spectrum was a feature used by the dolphin, the dolphin's errors were examined alongside similarities between the entire echo spectra of objects within each set. If the dolphin were using the shape of the spectrum, it would be expected to confuse objects with similar spectra. The echo spectra from object echoes at the 0 degree aspect (the aspect of the object that faced the front of the echoic choice apparatus) were normalized such that each individual spectrum had a maximum power of 1.0, and averaged over 10 measurement runs per object. These average spectra for each object were correlated with each other object within an object set. Object spectra were not correlated with objects outside their object set because the dolphin was asked to match objects within sets but not between sets. If the dolphin was using the shape of the spectra, we would expect the two objects that were the most often confused to have the highest r value. The r values for the correlations are shown in Table II alongside the proportion of errors made by the dolphin in each error type.

For four of the nine object sets, the two objects with the highest correlation between echo spectra were the two objects most often confused by the dolphin. For example, in the Green Foam set, the dolphin's predominant errors were between Slices and Wedges (63% of the errors). Slices and Wedges were more highly correlated ($r=0.952$) than either Slices and Mounds ($r=0.670$) or Mounds and Wedges ($r=0.635$). The Wood Plaques are considered to be one of these four sets because all three objects are highly correlated and all three objects are confused by the dolphin. For the other five object sets, the two objects with the highest correlation between echo spectra were *not* the two objects most often confused by the dolphin. For example, in the Stone Squares set, the dolphin's predominant errors were between Medium and Large (90% of the errors). However, Small and Large were more highly correlated ($r=0.787$) than Medium and Large ($r=0.425$) or Small and Medium ($r=0.365$).

TABLE II. Spectral correlations and the dolphin's errors.

Object set	Error type	Correlations r values	Errors Proportions
Strainers	Small-Medium	0.570	0.142
	Medium-Large	**0.708**	**0.786**
	Small-Large	0.580	0.071
Stone squares	Small-Medium	0.365	0.100
	Medium-Large	0.425	**0.900**
	Small-Large	**0.787**	0
Stone shapes	Cross-Line	0.858	**0.667**
	Line-Stack	0.856	0.333
	Cross-Stack	**0.952**	0
Foam cones	Diamond-Bow	**0.992**	**0.625**
	Trap-Diamond	0.914	0.250
	Trap-Bow	0.920	0.125
Figure 8's	Rope-Tubing	0.722	**0.833**
	Copper-Rope	0.698	0.083
	Copper-Tubing	**0.766**	0.083
Rods	Metal-Plastic	0.439	**0.364**
	Plastic-Wood	0.811	**0.318**
	Metal-Wood	**0.858**	**0.318**
Wood plaques	Holed-Plain	**0.995**	**0.280**
	Striped-Plain	**0.997**	**0.400**
	Hole-Striped	**0.999**	**0.320**
Sockets	Closed-Open	0.838	**0.933**
	Closed-Flat	**0.948**	0.067
	Open-Flat	0.906	0
Green foam	Slices-Mounds	0.670	0.375
	Slices-Wedges	**0.952**	**0.625**
	Mounds-Wedges	0.635	0

Note: The dolphin's predominant errors are highlighted in bold. When the dolphin did not make one type of error predominantly (Rods, Wood), all errors are highlighted. The object pairs with the highest correlation are also highlighted in bold. Object sets in which the highest correlation matches the predominant confusion are shown in bold and italics.

3. Linear model of the dolphin's performance

To test the hypothesis that the dolphin used a linear combination of multiple acoustic features, the dolphin's<> choice accuracy was compared to the classification accuracy of a statistical pattern classification model: the discriminant function analysis (DFA). In a DFA, one or more continuous predictor variables are used to form a linear model, from which stimulus classification is made. A separate DFA was run for each object set. The DFA was used to predict object type given the six single-echo acoustic features. For each object within a set, there were 230 echoes (23 object orientations $\times$ 10 measurement runs).

To compare the DFA results with the dolphin's classification results (12 trials per object in each set), the classification percentages for the original DFA counts were multiplied by 12. The classification matrices for the dolphin are presented above the transformed classification matrices for the DFA for three of the object sets in Table III. The DFA models were more accurate than the dolphin in overall matching accuracy, with the exception of the Green Foam object set. The classification performance of the DFA was compared against the dolphin's performance using chi square tests with the dolphin's choices as the expected distribution. The pattern

TABLE III. Classification matrices for the dolphin vs the DFA.

STONE SQUARES				
Dolphin's Performance		Choice		
		Small	Medium	Large
	Small	12	0	0
Sample	Medium	1	2	9
	Large	0	0	12
DFA Echo Classification		Choice		
		Small	Medium	Large
	Small	10.2	1.8	0.0
Sample	Medium	1.9	10.1	0.0
	Large	0.4	1.0	10.6
FIGURE 8'S				
Dolphin's Performance		Choice		
		Copper	Rope	Tubing
	Copper	12	0	0
Sample	Rope	1	5	6
	Tubing	1	4	7
DFA Echo Classification		Choice		
		Copper	Rope	Tubing
	Copper	9.9	0.0	2.1
Sample	Rope	0.0	11.5	0.5
	Tubing	0.0	1.4	10.6
SOCKETS				
Dolphin's Performance		Choice		
		Closed	Flat	Open
	Closed	6	0	6
Sample	Flat	1	11	0
	Open	8	0	4
DFA Echo Classification		Choice		
		Closed	Flat	Open
	Closed	10.7	0.7	0.6
Sample	Flat	0.9	10.8	0.3
	Open	1.5	0.9	9.5

of classification made by the DFA was significantly different from the dolphin for all of the sets except the Stone Shapes, Foam Cones, and the Green Foam (Stone Squares $[\chi^2(4)=42.86, N=36, p<0.001]$, Strainers $[\chi^2(4)=68.73, N=36, p<0.001]$, Stone Shapes $[\chi^2(4)=6.34, N=36, p>0.05]$, Foam Cones $[\chi^2(4)=3.87, N=36, p>0.05]$, Figure 8's $[\chi^2(4)=19.53, N=36, p<0.001]$, Rods $[\chi^2(4)=31.26, N=36, p<0.001]$, Wood Plaques $[\chi^2(4)=21.62, N=36, p<0.001]$, Sockets $[\chi^2(4)=21.63, N=36, p<0.001]$, Green Foam $[\chi^2(4)=2.98, N=36, p>0.05]$).

The DFA models did not produce the same biased classification patterns made by the dolphin for several object sets. For example, in both the Stone Squares set and the Strainers set, when the dolphin was presented with the Medium object he consistently chose the Large object. The DFA model did not show the same choice bias with the Medium object, it correctly classified the Medium object on most trials. For other object sets, the DFA did not replicate the pattern of errors made by the dolphin. For example, in the Figure 8's set, the dolphin's errors consisted primarily of confusions between Rope and Tubing (10/12 errors). In contrast, the DFA rarely misclassified Rope and Tubing. Another example is the Sockets set. The majority of the dolphin's errors were confusions between the Open Socket and the Closed Socket (14/15 errors). The DFA rarely misclassified the Open and Closed sockets.

When the classification performance of the DFA and the dolphin's performance were compared using values of d', the results show that the performance of the DFA and the dolphin were not similar. For two of the object sets (Rods and Wood), the DFA was more accurate in matching all of the object pairs (DFA's d' values higher than the dolphin's). For two of the object sets (Stone Shapes and Green Foam), the DFA was less accurate than the dolphin in matching all of the object pairs. The DFA's d' value was subtracted from the dolphin's d' value for each of the 27 object pairings (3 pairs $\times$ 9 object sets) to yield a difference score (e.g., for Small Stone vs Medium Stone: dolphin $d'=3.53$ vs DFA $d'=2.03$ for a difference score of 1.5). When the difference scores were between -0.3 and 0.3, the d' values were considered to be similar (scores ranged from -3.34 to 4.98). Out of the 27 object pairings, the d' values for the DFA were similar to the d' values for the dolphin for only 4 object pairings [Bow-Diamond (-0.2), Diamond-Trapezoid (0.29), Copper-Tubing (0.21), Metal-Plastic (-0.25)].

IV. GENERAL DISCUSSION

The present study explored two hypotheses regarding the way dolphins use acoustic information in echoes: (1) use of a single feature, or (2) use of a linear combination of multiple features. The results suggested that dolphins do not use a single feature across all object sets or a linear combination of the individual echo features. Five features appeared to be important to the dolphin on four or more object sets: the shape of the echo spectrum, the pattern of changes in target strength and number of highlights as a function of object orientation, and peak and center frequency. In eight of the nine object sets, the dolphin's error patterns implied the use of multiple acoustic features. Thus far, researchers have assumed that dolphins made use of multiple features in their echoes (Au, 1993), and these data provide some evidence for that assumption.

The first hypothesis included two possibilities: (A) use of a single, stable acoustic feature in all circumstances, and (B) use of a single acoustic feature, but the feature shifts based on the characteristics of the objects. For example, in case A, the dolphin always uses target strength regardless of whether objects vary in size, shape, or material, and in case B the dolphin uses peak frequency to discriminate between objects that vary in material and target strength to discriminate between objects that vary in size. The results of this study show that the dolphin did not use a single acoustic feature for all object sets, which eliminates case A. These data cannot completely rule out case B: the possibility that a dolphin may use a single, shiftable feature to discriminate among the objects (see Table I). However, the Stone Square set results argue against the use of a single feature since there was no single feature that could explain the dolphin's errors across all 12 measured features. The single feature hypothesis is improbable because there is no simple one-to-one relationship between the physical features of objects (e.g., size, shape, material) and the acoustic features of echoes (e.g., amplitude, frequency). For example, echo amplitude does

not map only to the size of the object, it is affected by the both the size and material of the object (along with other characteristics).

It would be more advantageous for a dolphin to use multiple acoustic features in the echoes than to use a single acoustic feature. One big advantage to using multiple features is that the values of some features would disambiguate other features. For example, if two objects had very similar target strengths, but different highlight structures, a dolphin attending to more than one feature would be able to discriminate between those objects. In many of the object sets analyzed in this study, two objects had similar values for one or more acoustic features yet the dolphin was able to discriminate between them.

Using certain acoustic features to disambiguate other features becomes particularly important when considering the variation in features as a function of orientation. This variation in features as a function of orientation was at times greater than the between-object variation! A major finding in this study was that *all* the measured acoustic features for nearly all the objects varied as a function of orientation. This means the echoes from an object in one orientation are quite different from the echoes from the *same* object ensonified from a different orientation. Two objects could have the same value for an acoustic feature from one orientation, but different values of that same feature from another orientation. A dolphin has the best chance of picking up differences between objects if it attends to multiple features, and gets multiple "looks" at an object from different orientations (this would correspond to the dolphin moving its head or swimming by the object).

The fact that the acoustic features of objects can vary so much with orientation actually creates a potentially salient "feature" for the dolphin: the pattern of changes in the echo across several orientations of the object. In a previous study, it appeared that the dolphin attended to this feature to discriminate between objects that varied in shape (Nachtigall *et al.*, 1980). The present results indicate that the dolphin could have used the pattern of changes in the echo (target strength bumpiness and highlight bumpiness) in six of the nine object sets. Future studies should measure object echoes from various orientations and consider the strong possibility that the dolphin gathers and uses acoustic information from more than one orientation (as long as the dolphin's head is unrestrained as it was in this study).

The hypothesis that dolphins use multiple acoustic features was tested by comparing the dolphin's performance to a discriminant function analysis (DFA). The dolphin's classification patterns and accuracy did not match the DFA's, which suggests that the dolphin may not have used a linear combination of these particular features to discriminate among the objects. One interpretation of these data is that the dolphin does in fact use a linear combination of features, but the features extracted in this study were incorrect and/or insufficient. In the current study we extracted features from the time waveforms and the spectra. It is also possible to extract spectrogram features or other time/frequency derived features (e.g., Okimoto *et al.*, 1998).

Another possibility is that the dolphin used a combination of multiple features that was complex and nonlinear. It is very likely that dolphins combine echo features nonlinearly to recognize objects. Human cognition is considered to be nonlinear in many ways, particularly when it comes to pattern recognition and decision-making. Because artificial neural networks resemble the organization of biological neural systems, they have been used frequently to model the performance of biological systems. There have been several efforts to model the echolocation performance of dolphins using neural networks (e.g., Au *et al.*, 1995; Roitblat *et al.*, 1989). The networks in these studies used temporal and/or spectral features of the echoes and performed at times better or worse than the dolphin. The drawback of these studies is that they did not attempt to compare the error patterns of the dolphin and the network, so it is difficult to ascertain whether the network was using the same features as the dolphin. Since neural networks are capable of performing nonlinearly, this is a promising method of investigating how dolphins use acoustic features if the errors patterns of the dolphin and the network are analyzed.

It is a limitation of the current study that the sonar signals from the actual dolphin and the resulting echoes were not recorded and measured during the behavioral task. The acoustic features of the echoes are affected by not only the orientation of the objects, but also the characteristics of the dolphin's outgoing sonar signal. The objects in this study were measured in a test tank using a dolphin click that did not vary from trial to trial. In contrast, dolphin signals can vary in both frequency and intensity (Au, 1993). An individual dolphin's signals varied in peak frequency from 60 to 140 kHz (Au, 1980). Dolphins can vary the amplitude of their signals over a large dynamic range (e.g., peak-to-peak source levels vary from 150–230 dB re 1 μPa; Au, 1993). Dolphins modify their signals in an adaptive manner according to the environment and task demands. For example, dolphins will emit higher amplitude signals in noisy environments (Au *et al.*, 1974). In a situation where a dolphin is performing a specific task over and over again, there can be fluctuations in the characteristics of the emitted clicks during the task (Au, 1993). These variations in a dolphin's outgoing clicks will influence the acoustic characteristics of the resulting object echoes. For example, the amplitude of the echoes from an object is dependent on the source level of the projected click (Au, 1993).

To obtain the most accurate perspective on how dolphins use acoustic information in echoes, the actual outgoing signals and echoes should be recorded during the task. This was not done during the current study because it would have added additional animal training time that was not feasible given time constraints. To record the dolphin's signals in a free-swimming task such as this one, the dolphin must be trained to perform the matching task efficiently while carrying a bite-plate hydrophone. This difficult training procedure can take many months to a year. Because of these constraints, Toby's click signals were not recorded. It is assumed that he was using typical clicks resembling the recorded click that was used to obtain echoes for these acoustic measurements.

In the present study, statistical methods were used to determine between-object differences in echo features. A potential problem with this method is that statistical significance may not equal biological significance. Dolphins' ability to discriminate between sounds with different intensities is known to be approximately 1 dB (Vel'min, Titov, and Yurkevich, 1975, cited in Au, 1993). In the Figure 8's set, Rope has a target strength of −32.9 dB, which is *statistically* significantly different than the target strength for Tubing (−32.3 dB). However, dolphins may not perceive objects with target strengths within 1 dB of each other to be different since 1 dB is about the limit of their intensity discrimination ability. The Figure 8's set is the only object set in which objects that are separated by less than 1 dB are statistically significantly different. In all other sets, objects that are statistically different should also be perceived as different to the dolphin (i.e., objects are separated by 1 dB or more).

Dolphins' ability to discriminate between differences in the duration, frequency, bandwidth, or number of highlights in short broadband echoes is unknown. This makes it difficult to compare statistical and biological significance for these features. For example, the peak frequency of the Small Strainer (119.1 kHz) was statistically significantly different than the Medium (113.6 kHz) and Large Strainer (114.6 kHz). Since dolphins' frequency discrimination ability for click signals has not been comprehensively assessed, it is unknown whether the dolphin could actually discriminate between an echo with a peak frequency of 114 kHz and an echo with a peak frequency of 119 kHz. Dolphins can discriminate between a 114.0 kHz tone and a 114.2 kHz tone (Herman and Arbeit, 1972). Assuming that dolphins' frequency discrimination ability for click signals is comparable to their ability to discriminate between tonal signals of different frequencies, it suggests that dolphins can discriminate between an echo with a peak frequency of 114 kHz and an echo with a peak frequency of 118 kHz. One study found that dolphins can detect broadband signals slightly better than a pure-tone signal, but ability to discriminate between broadband sounds with different peak or center frequencies was not assessed (Au *et al.*, 2002). Assuming that dolphins' ability to discriminate temporal differences in click signals is comparable to their ability for tonal signals (Yunker and Herman, 1974), a dolphin could tell the difference between an echo that is 100 and 109 μs (using a difference limen of 0.06). Clearly, studies of dolphins' ability to discriminate between differences in the duration, bandwidth, or number of highlights in echoes are needed to make further progress on determining the salient features of echoes using the error analysis method presented here.

ACKNOWLEDGMENTS

We would like to thank Kelly Benoit-Bird, Aran Mooney, and Kim Andrews for their assistance in making the acoustic measurements of the targets. We are grateful to the trainers and staff at Epcot's Living Seas, particularly E. Putman, A. Stamper, D. Bickel, D. Clark, J. Gory, W. Fellner, M. Muraco, M. Barringer, B. Cavanaugh, L. Davis, D. Feuerbach, C. Goonen, L. Larsen-Plott, K. Odell, C. Litz, J. Davis, T. Hopkins, B. Stevens, J. Mellen, and J. Ogden. We thank Paul Nachtigall, Patricia Couvillon, Ann Peters, Nathaniel Gibbs, Kelly Benoit-Bird, and James Simmons for their constructive suggestions on earlier versions of this manuscript. Funding for this project was provided in part by grants to CMD from the American Psychological Association (Dissertation Research Award), American Association of University Women (Pacific Dissertation Fellowship), Soroptimist International (Founder Region Dissertation Fellowship), and SEASPACE (Scholarship Award) and to HEH from the Walt Disney Company's Animal Programs and New College of Florida. This publication is contribution number 1206 from the Hawaii Institute of Marine Biology.

[1]The value d' is usually calculated for a detection task, but can also be calculated for a discrimination task (match-to-sample) in which the participant is presented with a sample object and must choose a match from among three alternatives. Data collected in this three-alternative procedure for each object set were broken down into three 2×2 confusion matrices to calculate d'. In the following example, the 2×2 matrix includes only objects 1 and 2 (trials with object 3 are not counted). A hit is the choice of object 1 when object 1 was the correct match, and a false alarm is the choice of object 1 when object 2 was the correct match. d' was calculated using the proportion of hits (the number of trials in which the dolphin correctly responded that object 1 was the sample, divided by the total number of trials in which 1 was the sample) vs the proportion of false alarms (the number of trials in which the dolphin responded that object 1 was the sample, divided by the total number of trials in which object 2 was the sample). A d' value of zero indicates no discrimination behavior, and is comparable to 50% correct on a two-alternative task. A d' value of 1.0 is comparable to approximately 76% correct responses on a two-choice task.

Au, W. W. L. (**1980**). "Echolocation signals of the Atlantic bottlenose dolphin (*Tursiops truncatus*) in open waters," in *Animal Sonar Systems*, edited by R. G. Busnel and J. F. Fish (Plenum Press, New York), pp. 251–282.

Au, W. W. L. (**1993**). *The Sonar of Dolphins* (Springer, New York).

Au, W. W. L. (**2000**). "Echolocation in dolphins," in *Hearing by Whales and Dolphins*, edited by W. W. L. Au, A. N. Popper, and R. R. Fay (Springer New York), pp. 364–408.

Au, W. W. L., Andersen, L. N., Rasmussen, A. R., Roitblat, H. L., and Nachtigall, P. E. (**1995**). "Neural network modeling of a dolphin's sonar discrimination capabilities," J. Acoust. Soc. Am. **98**, 43–50.

Au, W. W. L., Floyd, R. W., Penner, R. H., and Murchison, A. E. (**1974**). "Measurement of echolocation signals of the Atlantic bottlenose dolphin, *Tursiops truncatus* Montagu, in open waters," J. Acoust. Soc. Am. **54**, 1280–1290.

Au, W. W. L., Lemonds, D. W., Vlachos, S., Nachtigall, P. E., and Roitblat, H. L. (**2002**). "Atlantic bottlenose dolphin (*Tursiops truncatus*) hearing thresholds for brief broadband signals," J. Comp. Psychol. **116**, 151–157.

Au, W. W. L., and Martin, D. (**1988**). "Sonar discrimination of metallic plates by dolphins and humans," in *Animal Sonar: Processes and Performance*, edited by P. E. Nachtigall and P. W. B. Moore (Plenum Press, New York), pp. 809–813.

Au, W. W. L., and Pawloski, J. L. (**1992**). "Cylinder wall thickness difference discrimination by an echolocating Atlantic bottlenose dolphin," J. Comp. Physiol., A **172**, 41–47.

Au, W. W. L., and Turl, C. W. (**1991**). "Material composition discrimination of cylinders at different aspect angles by an echolocating dolphin," J. Acoust. Soc. Am. **89**, 2448–2451.

Bauer, G. B., and Johnson, C. M. (**1994**). "Trained motor imitation by bottlenose dolphins (*Tursiops truncatus*)," Percept. Mot. Skills **79**, 1307–1315.

Benoit-Bird, K. J., Au, W. W. L., and Kelley, C. D. (**2003**). "Acoustic backscattering by Hawaiian lutjanid snappers. I. Target strength and swimbladder characteristics," J. Acoust. Soc. Am. **114**, 2757–2766.

Green, D. M., and Swets, J. A. (**1988**). *Signal Detection Theory and Psychophysics* (Peninsula, Los Altos, CA).

Hammer, C. E., Jr., and Au, W. W. L. (**1980**). "Porpoise echo-recognition:

An analysis of controlling target characteristics," J. Acoust. Soc. Am. **68**, 1285–1293.
Harley, H. E., Putman, E. A., and Roitblat, H. L. (**2003**). "Bottlenose dolphins perceive object features through echolocation," Nature (London) **424**, 667–669.
Herman, L. M., and Arbeit, W. R. (**1972**). "Frequency discrimination limens in the bottlenose dolphin: 1-70Ks/c," J. Auditory Res. **2**, 109–120.
McClellan, M. E., and Small, A. M. (**1965**). "Time-separation pitch associated with correlated noise burst," J. Acoust. Soc. Am. **38**, 142–143.
Nachtigall, P. E., Murchison, A. E., and Au, W. W. L. (**1980**). "Cylinder and cube discrimination by an echolocating blindfolded bottlenose dolphin," in *Animal Sonar Systems*, edited by R. G. Busnel and J. F. Fish (Plenum Press, New York), pp. 945–947.
Okimoto, G., Shizumura, R., and Lemonds, D. (**1998**). "Active biosonar systems based on multiscale signal representations and hierarchical neural networks," in *Detection and remediation technologies for mines and minelike targets III*, Proc. SPIE **3392**, 316–323.
Roitblat, H. L., Moore, P. W. B., Nachtigall, P. E., Penner, R. H., and Au, W. W. L. (**1989**). "Natural echolocation with an artificial neural network," Int. J. Neural Networks **1**, 239–248.
Small, A. M. and McClellan, M. E. (**1963**). "Pitch associated with time delay between two pulse trains," J. Acoust. Soc. Am. **35**, 1246–1255.
Xitco, M. J., Jr. (1996). "Referential pointing by bottlenose dolphins," Southern Methodist University, Dallas (unpublished doctoral dissertation).
Xitco, M. J., Jr., and Roitblat, H. L. (**1996**). "Object recognition through eavesdropping: Passive echolocation in bottlenose dolphins," Anim. Learn Behav. **24**, 355–365.
Yunker, M. P., and Herman, L. M. (**1974**). "Discrimination of auditory temporal differences by the bottlenose dolphin and by the human," J. Acoust. Soc. Am. **56**, 1870–1875.

Elastic stiffness coefficients (c_{11}, c_{33}, and c_{13}) for freshly excised and formalin-fixed myocardium from ultrasonic velocity measurements

Min Yang, Steven L. Baldwin, Karen R. Marutyan, Kirk D. Wallace, Mark R. Holland, and James G. Miller[a)]

Department of Physics, Washington University, One Brookings Drive, Campus Box 1105, St. Louis, Missouri 63130

(Received 10 September 2005; revised 4 January 2006; accepted 5 January 2006)

The goal of this study was to measure elastic stiffness coefficients of freshly excised and subsequently formalin-fixed myocardial tissue. Our approach was to measure the angle-dependent phase velocities associated with the propagation of a longitudinal ultrasonic wave (3–8 MHz) in ovine myocardium using phase spectroscopy techniques and to interpret the results in the context of orthotropic and transversely isotropic models describing the elastic properties of myocardium. The phase velocity results together with density measurements were used to obtain the elastic stiffness coefficients c_{11}, c_{33}, and c_{13} for both symmetries. The results for the elastic stiffness coefficients c_{11}, c_{33}, and c_{13} are the same for both symmetries. Measurements for freshly excised myocardium and the same tissue after a period of formalin fixation were compared to examine the impact of fixation on the elastic stiffness coefficients. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2168547]

PACS number(s): 43.80.Cs, 43.80.Ev [CCC] Pages: 1880–1887

I. INTRODUCTION

Stress and strain distributions in the heart wall may be of value in the study of regional ventricular function. Recently introduced extensions of echocardiography permit the evaluation of regional myocardial deformation through strain and strain-rate imaging (D'Hooge *et al.*, 2000, 2002; Eroglu *et al.*, 2005; Gilman *et al.*, 2004; Inaba *et al.*, 2005; Sutherland *et al.*, 2004; Weidemann *et al.*, 2003; Williams *et al.*, 2005; Yuda *et al.*, 2002). Because of the relationship with stress and strain, characterization of the elastic moduli of myocardium may therefore be of interest.

Approaches using strain and strain-rate imaging employed for regional cardiac performance often assume that the heart is an isotropic medium. However, myocardium consists of myocytes that are arranged into locally parallel muscle fibers, which in turn are embedded in an extracellular matrix (Humphery, 2001; Streeter and Hanna, 1973). This results in the ultrasonic properties of the heart exhibiting a dependence on angle of insonification relative to the direction of the myofibers (Baldwin *et al.*, 2005a, b; Hoffmeister *et al.*, 1996a; Mottley and Miller, 1990; Verdonk *et al.*, 1996). We and others have modeled myocardium as having transversely isotropic symmetry because of the locally uniaxial architecture of the myofibers (Hoffmeister *et al.*, 1995, 1996b; O'Donnell and Skovoroda, 2004; Verdonk *et al.*, 1992). The elastic properties of a medium that exhibits this symmetry can be described by five distinct stiffness coefficients. However, a careful study by LeGrice *et al.* (1995) suggests that ventricular myocardium has a discrete laminar (sheetlike) organization that exhibits a lower symmetry than that of a uniaxial structure. This observation may play a further role in understanding mechanical properties of myocardium. For these reasons, an orthotropic material model for myocardium has been suggested (Usyk *et al.*, 2000). In the study by Usyk *et al.*, it is further suggested that, compared with the transversely isotropic model, the orthotropic model of passive myocardial properties yields improved agreement with measured end-diastolic strains provided that the stiffness transverse to the myocardial laminate is lower than that within the sheet (Usyk *et al.*, 2000). With orthotropic symmetry, a total of nine distinct stiffness coefficients are required to describe the elastic properties.

The objective of the current study was to determine the elastic stiffness coefficients c_{11}, c_{33}, and c_{13} in myocardium in the context of both orthotropic and transversely isotropic symmetries from measurements of ultrasonic phase velocity (3–8 MHz) and density. Previously our laboratory reported time-of-flight measurements of velocity on freshly excised and formalin-fixed ovine myocardium (Baldwin *et al.*, 2005b). In contrast, the current study made use of a phase spectroscopy technique to analyze the same set of radiofrequency data to obtain ultrasonic phase velocity at specific frequencies between 3 and 8 MHz. The impact of formalin fixation on the elastic stiffness coefficients was also examined.

II. METHODS

A. Measurements of ultrasonic phase velocity in myocardium

The experimental procedure used for velocity measurements in the current study has been reported previously by our laboratory (Baldwin *et al.*, 2005b) and is summarized

[a)]Author to whom correspondence should be addressed. Electronic mail: james.g.miller@wustl.edu

 0001-4966/2006/119(3)/1880/8/$22.50

below. A total of 36 cylindrical specimens were cored from 12 freshly excised ovine hearts. The hearts were from a commercial slaughterhouse within 30 min of the death of the lambs (about 6–7 months of age). Specimens were obtained from three specific locations within each heart (two from the left ventricular free wall and one from the septum). Broadband (3–8 MHz) through-transmission radiofrequency-based measurements (Panametrics 5800 Pulser/Receiver, Panametrics, Waltham, MA) were performed in a water-filled tank at room temperature with a matched pair of focused piezoelectric transducers (Panametrics V309, $\frac{1}{2}$-in. diameter, 2-in. focal length, 5-MHz center frequency, Panametrics, Waltham, MA). Data were acquired on freshly excised specimens as a function of the angle of insonification relative to the predominant orientation of the myofibers in increments of 5° over a complete rotation using a digital oscilloscope (Tektronix 2430A, Tektronix, Beaverton, OR). A previous study from our laboratory (Verdonk *et al.*, 1992) has shown the transmural shift in fiber orientation across the ventricular wall. However, in this current study, care was taken to make sure that the ultrasonic beam (beam diameter approximately 1–2 mm) was centered on the midmyocardial region of the tissue specimens where the myofibers are locally uniaxially aligned. In addition, immediately after the acquisition of through-transmission data, corresponding pulse-echo data were acquired for specimen thickness measurements at each incremented angle. Corresponding water path reference data were obtained at a temperature within 0.1 °C of that for the corresponding specimen data. An identical set of measurements was later repeated on the same tissue specimens that had been stored in formalin for several months in order to provide a sufficient amount of time to ensure that the process of fixation had reached completion.

Phase spectroscopy was used to determine the phase velocity in the specimens. In this approach, the speed of sound in the specimen (V_{specimen}) at angle θ for a given frequency ω is obtained by

$$V_{\text{specimen}}(\omega,\theta)=\frac{V_{\text{water}}}{1+[V_{\text{water}}/L(\theta)][\Delta\varphi(\omega,\theta)/\omega]}, \tag{1}$$

where V_{water} is the speed of sound in water, $L(\theta)$ is the thickness of the myocardial specimen at angle of rotation θ, and $\Delta\varphi(\omega,\theta)=\varphi_{\text{specimen}}(\omega,\theta)-\varphi_{\text{water}}(\omega)$ denotes the difference in phases between the through-specimen and water reference radio frequency pulses. The speed of sound in water was estimated based on a fifth-order polynomial expression as a function of temperature (Del Grosso and Mader, 1972). To calculate the phase, radio frequency (rf) signals were symmetrized to avoid excess phase wrapping (Trousil *et al.*, 2001). Time-of-flight broadband velocity results (Baldwin *et al.*, 2005b) were chosen to resolve phase sheet ambiguity. As a result of the symmetry of our cylindrical specimens, the angle of insonification relative to the predominant myofiber orientation ranged from 0° to 90° a total of four times over a complete rotation. This permits a fourfold averaging of the phase velocity data at each angle between 0° and 90° relative to the myofiber orientation.

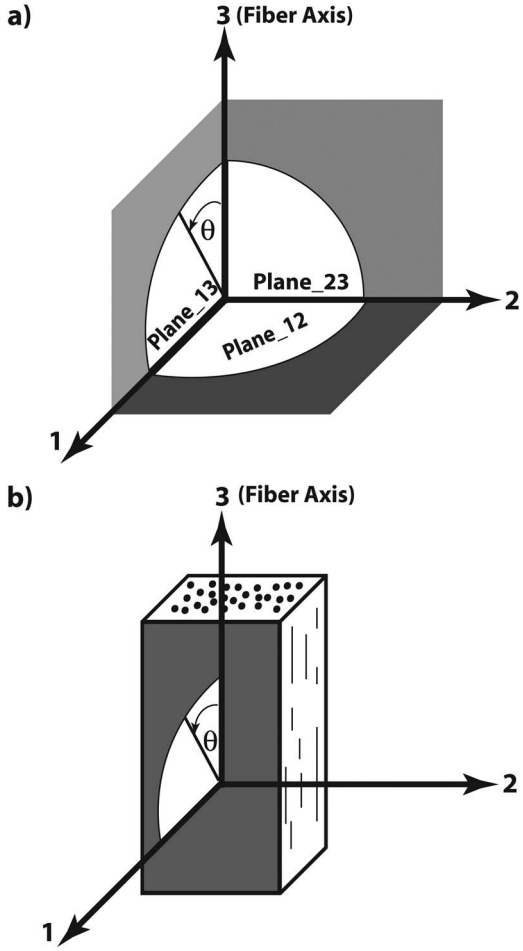


FIG. 1. (a) Coordinate axes and planes with mirror symmetry in an orthotropic medium. (b) Coordinate axes in a transversely isotropic medium.

B. Density measurements

Density measurements were performed on freshly excised and formalin-fixed specimens from representative hearts. Based on Archimedes' principle, the density was determined from the actual weight (w_{actual}) and the apparent weight of the specimen submerged in water ($w_{\text{submerged}}$) according to the equation

$$\rho_{\text{specimen}}=\rho_{\text{water}}/(1-w_{\text{submerged}}/w_{\text{actual}}). \tag{2}$$

The actual weight of each density specimen was measured using an electronic analytical balance (Fisher Scientific Model 2200). A higher precision automatic electrobalance (Cahn model 29) was used to suspend the specimen in water and measure the submerged weight. The density of water was taken to be 0.998 g/cm^3 (the density of water at 20 °C). The results for all specimens were averaged and the standard deviation was calculated.

C. Determining stiffness coefficients from the phase velocity

A brief review of linear elastic theory under the symmetries examined in this investigation is provided in the Appendix. For orthotropic symmetry, in which there exist three orthogonal planes of mirror symmetry [Plane_12, Plane_13 and Plane_23 as shown in Fig. 1(a)], nine distinct coefficients

TABLE I. The relationships of stiffness coefficients among different degrees of symmetry.

Orthotropic:	c_{11}^{Ortho} c_{22}^{Ortho} c_{33}^{Ortho}	c_{12}^{Ortho} c_{13}^{Ortho} c_{23}^{Ortho}	c_{44}^{Ortho} c_{55}^{Ortho} c_{66}^{Ortho}
Transversely Isotropic:	c_{11}^{TIso} c_{33}^{TIso}	$c_{11}^{TIso}-2c_{66}^{TIso}$ c_{13}^{TIso}	c_{44}^{TIso} c_{66}^{TIso}
Isotropic:	c_{11}^{Iso}	$c_{11}^{Iso}-2c_{44}^{Iso}$	c_{44}^{Iso}

are required to describe the elastic properties of a medium. A transversely isotropic medium exhibits a higher symmetry. This is illustrated in Fig. 1(b) which depicts a Cartesian coordinate system in which the 3-axis represents the one distinct direction and the other two axes (1 and 2) are equivalent to each other. The elastic properties of the medium can then be described with five distinct coefficients. In the case of highest symmetry, that of isotropy, the elastic matrix contains only two independent coefficients. The relationships among the stiffness coefficients for different degrees of symmetry are summarized in Table I.

The phase velocities for ultrasonic waves propagating in a specific medium can be found using the Christoffel equation (Auld, 1990). For orthotropic symmetry, the expression for phase velocity of a quasi-longitudinal mode wave propagating in Plane_13 is

$$V_{\text{phase\_ql}}^{\text{Ortho}}(\theta)=\sqrt{\frac{A^{\text{Ortho}}(\theta)+\sqrt{[B^{\text{Ortho}}(\theta)]^2+4C^{\text{Ortho}}(\theta)}}{2\rho}}, \tag{3}$$

where

$$A^{\text{Ortho}}(\theta)=c_{11}^{\text{Ortho}}\sin^2\theta+c_{33}^{\text{Ortho}}\cos^2\theta+c_{55}^{\text{Ortho}},$$

$$B^{\text{Ortho}}(\theta)=(c_{11}^{\text{Ortho}}-c_{55}^{\text{Ortho}})\sin^2\theta-(c_{33}^{\text{Ortho}}-c_{55}^{\text{Ortho}})\cos^2\theta,$$

$$C^{\text{Ortho}}(\theta)=(c_{13}^{\text{Ortho}}+c_{55}^{\text{Ortho}})^2\sin^2\theta\cos^2\theta,$$

ρ is the density of the medium, and θ is the angle of insonification relative to the 3-axis. For transversely isotropic symmetry, the phase velocity for a wave propagating in Plane_13 takes a similar form

$$V_{\text{phase\_ql}}^{\text{TIso}}(\theta)=\sqrt{\frac{A^{\text{TIso}}(\theta)+\sqrt{[B^{\text{TIso}}(\theta)]^2+4C^{\text{TIso}}(\theta)}}{2\rho}} \tag{4}$$

where

$$A^{\text{TIso}}(\theta)=c_{11}^{\text{TIso}}\sin^2\theta+c_{33}^{\text{TIso}}\cos^2\theta+c_{44}^{\text{TIso}},$$

$$B^{\text{TIso}}(\theta)=(c_{11}^{\text{TIso}}-c_{44}^{\text{TIso}})\sin^2\theta-(c_{33}^{\text{TIso}}-c_{44}^{\text{TIso}})\cos^2\theta,$$

$$C^{\text{TIso}}(\theta)=(c_{13}^{\text{TIso}}+c_{44}^{\text{TIso}})^2\sin^2\theta\cos^2\theta.$$

A comparison of Eqs. (3) and (4) shows that contributions from the stiffness coefficients to the angular dependence of velocity are almost identical except that the pure shear stiffness coefficient c_{55} is present for orthotropic symmetry whereas the coefficient c_{44} is present for transversely isotropic symmetry. It is important to note that under transversely isotropic symmetry, when the 3-axis is chosen as the symmetry axis, c_{55} is identical to c_{44}. Thus for the phase velocity of a quasi-longitudinal wave propagating in Plane_13, Eqs. (3) and (4) have the same form. For the isotropic case, the phase velocity is independent of direction, $V_{\text{longitudinal}}^{\text{Iso}}=\sqrt{c_{11}^{\text{Iso}}/\rho}$.

From Eqs. (3) and (4), it is seen that c_{11} and c_{33} for both the orthotropic and the transversely isotropic cases can be determined when the ultrasound is propagating perpendicular to the 3-axis ($\theta=\pi/2, 3\pi/2$) and parallel to the 3-axis ($\theta=0, \pi, 2\pi$), respectively, i.e.,

$$c_{11}=\rho V_{\text{perpendicular}}^2 \tag{5}$$

and

$$c_{33}=\rho V_{\text{parallel}}^2. \tag{6}$$

In this study, mean phase velocity values perpendicular (along the 1-axis direction) and parallel (along the 3-axis direction) to the myofiber direction, together with the density results, were used to calculate c_{11} and c_{33} at specific frequencies using Eqs. (5) and (6). Fractional uncertainties of c_{11}, c_{33} were estimated by

$$\frac{\Delta c_{11}}{c_{11}}=\sqrt{\left(\frac{\Delta\rho}{\rho}\right)^2+4\left(\frac{\Delta V_{\text{perpendicular}}}{V_{\text{perpendicular}}}\right)^2} \tag{7}$$

and

$$\frac{\Delta c_{33}}{c_{33}}=\sqrt{\left(\frac{\Delta\rho}{\rho}\right)^2+4\left(\frac{\Delta V_{\text{parallel}}}{V_{\text{parallel}}}\right)^2}. \tag{8}$$

However, the expressions for c_{13} in both cases, found by rearranging the terms in Eqs. (3) and (4), have relatively complicated forms that depend on shear elastic components. The elastic stiffness coefficient c_{13} for transversely isotropic symmetry is given by

$$c_{13}^{\text{TIso}}=[D^{\text{TIso}}(\theta)-A^{\text{TIso}}(\theta)\rho V_{\text{phase\_ql}}^{\text{TIso }2}+\rho^2 V_{\text{phase\_ql}}^{\text{TIso }4}]^{1/2}-c_{44}^{\text{TIso}}, \tag{9}$$

where $A^{\text{TIso}}(\theta)=c_{11}^{\text{TIso}}\sin^2\theta+c_{33}^{\text{TIso}}\cos^2\theta+c_{44}^{\text{TIso}}$ and

$$\begin{aligned}D^{\text{TIso}}(\theta)&=\left(\tfrac{1}{4}\right)[[A^{\text{TIso}}(\theta)]^2-[B^{\text{TIso}}(\theta)]^2]\\&=(c_{33}^{\text{TIso}}\cos^2\theta+c_{44}^{\text{TIso}}\sin^2\theta)(c_{11}^{\text{TIso}}\sin^2\theta\\&\quad+c_{44}^{\text{TIso}}\cos^2\theta).\end{aligned}$$

For orthotropic symmetry, c_{13} has the same form, with c_{55} in place of c_{44}. From Eq. (9), it is apparent that c_{13} depends on c_{55} for orthotropic symmetry and on c_{44} for transversely isotropic symmetry, where c_{55} and c_{44} correspond to pure shear wave propagation relative to the myofibers. Our laboratory has previously proposed a method

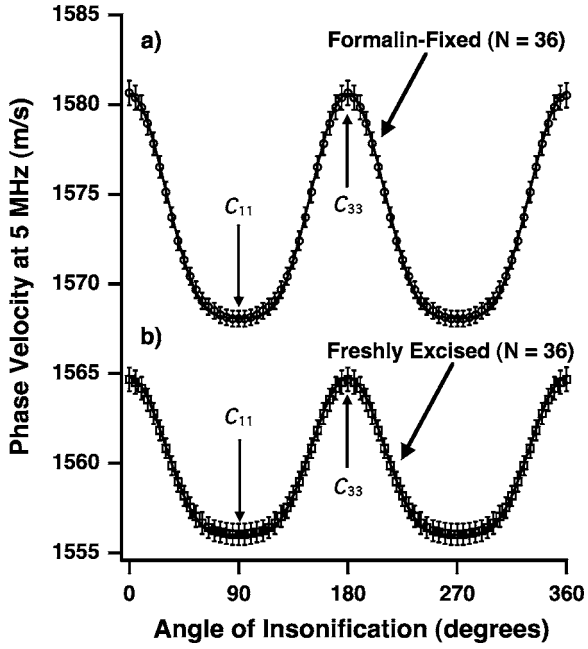


FIG. 2. Average measured phase velocity at 5 MHz (mean±standard error, $N=36$) as a function of angle of insonification for (a) formalin-fixed myocardium (open circles) and (b) freshly excised myocardium (open squares) with resulting curve fits (solid lines).

to estimate the stiffness coefficient c_{13} from the ultrasonic quasi-longitudinal velocities as a function of angle (Hoffmeister *et al.*, 1995). Following that approach, we inserted the measured values of c_{11}, c_{33}, and ρ into Eq. (3) [or Eq. (4)]. Then, based on physical grounds as will be discussed later, the shear elastic stiffness coefficient c_{55} (or c_{44}) was assumed to be much smaller than c_{11} and c_{33}. The velocity values calculated from Eq. (3) [or Eq. (4)] as a function of angle of insonification for varying c_{13} were compared with the experimentally measured phase velocity curve. The c_{13} value was determined to be that which provides the best fit to the experimental data. Specifically, a one-parameter user-defined function was created in Igor Pro (Wavemetrics, Lake Oswego, OR) and curve fits were performed to the measured velocities as a function of angle with this user-defined function to obtain c_{13}. The best fit is achieved when the sum of square difference (weighted by the standard deviation of measured data) between the fitted value for a given c_{13} and the measured data value reaches a minimum. The value of the shear coefficient was initially set to be zero and then varied over several orders of magnitude to evaluate its influence on the estimation of c_{13}. The best-fit result for c_{13} was obtained for each fixed value of c_{55} (or c_{44}). Here, the physical realizability conditions require that $c_{13}^2 < c_{11}c_{33}$.

III. RESULTS

The average phase velocity, from which the elastic stiffness coefficients were determined, exhibits a distinct directional dependence (anisotropy) with a minimum perpendicular to and a maximum parallel to the myofibers for all frequencies in the experimental bandwidth (3–8 MHz). Figure 2 shows the phase velocity at 5 MHz as a function of angle of insonification relative to the predominant orientation of the myofibers. The measured mass densities were 1.059 ± 0.002 g/cm^3 (mean±standard deviation) for freshly excised myocardium and 1.062 ± 0.002 g/cm^3 for formalin-fixed myocardium.

Tables II and III show the values of c_{11} and c_{33} for both orthotropic and transversely isotropic symmetries determined via Eqs. (5) and (6) along with the associated phase velocities at specific frequencies in the experimental bandwidth for both freshly excised and formalin-fixed myocardium. The ratios c_{33}/c_{11} and $(c_{33}-c_{11})/c_{33}$ are also given as a measure of the relative anisotropy in the elastic stiffness coefficients. The uncertainties of c_{11} and c_{33} were obtained from Eqs. (7) and (8). As an example, the uncertainties of elastic coefficients for freshly excised myocardium at 5 MHz are presented here. Using $\rho_{\text{fresh}}=1.059\pm0.002$ g/cm^3, $V_{\text{perpendicular}}=1556\pm4$ m/s, and $V_{\text{parallel}}=1565\pm4$ m/s, and propagating the uncertainties, one finds $c_{11}=2.564\pm0.015$ GPa and $c_{33}=2.593\pm0.014$ GPa. Figure 3 presents a direct comparison of the elastic stiffness coefficients (c_{11} and c_{33} at 5 MHz) for freshly excised and formalin-fixed myocardium. From the

TABLE II. Summary of results of c_{11}, c_{33} and the phase velocities at frequencies from 3 to 8 MHz for freshly excised myocardium. Four significant figures were kept for viewing and five significant figures were used for calculation. Representative values of standard deviations are shown in the last row.

Frequency (MHz)	Velocity parallel to fiber axis (m/s)	Velocity perpendicular to fiber axis (m/s)	c_{33} (GPa)	c_{11} (GPa)	$\frac{c_{33}}{c_{11}}$	$\frac{c_{33}-c_{11}}{c_{33}}$
3	1563	1554	2.586	2.562	1.009	0.009
4	1564	1556	2.590	2.563	1.011	0.011
5	1565	1556	2.593	2.564	1.011	0.011
6	1565	1556	2.595	2.565	1.012	0.012
7	1566	1557	2.597	2.566	1.012	0.012
8	1567	1557	2.599	2.566	1.013	0.013
		Representative standard deviations				
⋯	±4	±4	±0.014	±0.015	⋯	⋯

TABLE III. Summary of results of c_{11}, c_{33} and the phase velocities at frequencies from 3 to 8 MHz for formalin-fixed myocardium. Four significant figures were kept for viewing and five significant figures were used for calculation. Representative values of standard deviations are shown in the last row.

Frequency (MHz)	Velocity parallel to fiber axis (m/s)	Velocity perpendicular to fiber axis (m/s)	c_{33} (GPa)	c_{11} (GPa)	$\frac{c_{33}}{c_{11}}$	$\frac{c_{33}-c_{11}}{c_{33}}$
3	1579	1567	2.646	2.608	1.015	0.014
4	1580	1568	2.651	2.610	1.016	0.015
5	1581	1568	2.653	2.611	1.016	0.016
6	1581	1569	2.656	2.613	1.017	0.016
7	1582	1569	2.658	2.614	1.017	0.017
8	1582	1569	2.659	2.614	1.017	0.018
			Representative standard deviations			
⋯	±4	±3	±0.014	±0.011	⋯	⋯

results shown in Tables II and III, it is evident that formalin fixation stiffens the myocardial tissue, as observed qualitatively.

As indicated in Sec. II, for insonification angles in Plane_13, orthotropic and transversely isotropic symmetry exhibit an analogous representation of wave propagation characteristics. When the one-parameter fit technique is used to estimate c_{13}, the procedures are the same for both symmetry cases. The same results for c_{13} would be obtained. Hence, in what follows, we only refer to the transversely isotropic symmetry. The best-fit curve used to determine c_{13} showed good agreement with the average measured phase velocity data for all angles and frequencies for both the freshly excised and the formalin-fixed cases. A representative example for results at 5 MHz is displayed in Fig. 2. Estimated values of c_{13} determined from curve fitting, where c_{44} was held constant, are summarized in Tables IV and V for a wide range of specific values of c_{44}. Also shown in those two tables are corresponding shear wave velocities for the assumed c_{44} values. The influence of c_{44} on the determination of c_{13} is discussed in detail in the next section.

IV. DISCUSSION

As shown in the results section, the three elastic stiffness coefficients c_{11}, c_{33}, and c_{13} were estimated to be the same for both orthotropic and transversely isotropic symmetries. In the process of estimating c_{13}, we have made the assumption that c_{44} is negligibly small compared with c_{11} and c_{33}. The validity of our assumption for c_{44} is supported by previous studies. Frizzell and Carstensen (1976) showed that the stiffness coefficient is less than 1 MPa for fresh canine soft tissue from measurements of acoustic impedance at low megahertz frequencies (2–14 MHz). Madsen *et al.* (1983) implemented a similar technique to investigate the shear properties of fresh bovine soft tissue and obtained 0.124 MPa for the shear modulus of cardiac muscle. An earlier study from our laboratory (Hoffmeister *et al.*, 1996a) reported $c_{44}=8.97$ MPa for formalin-fixed human myocardium using the sampled-continuous wave (sampled-CW) technique with a 3-MHz transverse-mode transducer. In Table IV for freshly excised tissue, in the seemingly possible range of $c_{44}\leqslant 1$ MPa, c_{13} varied by less than 1%. Even when c_{44} was assumed to be two orders of magnitude higher than the upper limit of 1 MPa, the change in c_{13} was less than 10%. In the case of formalin-fixed myocardium, a larger, seemingly possible range was considered based on results from Hoffmeister *et al.* (1996a), as shown in Table V. For the larger range over which c_{44} was altered ($c_{44}\leqslant 10$ MPa), the changes in c_{13} were subtle (less than 1%). Even for the physically unreasonable value of 100 MPa, c_{13} showed no more than a 10% change.

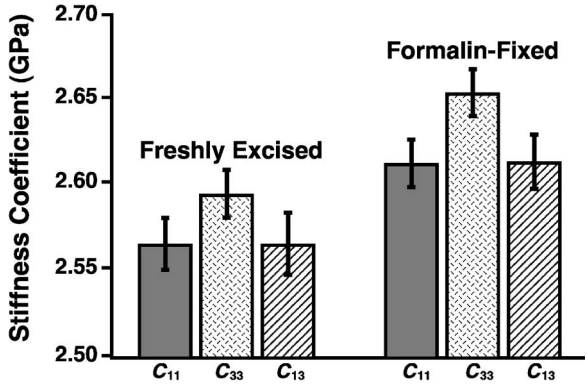


FIG. 3. Comparison of the elastic coefficients at 5 MHz for freshly excised and formalin-fixed myocardium. Standard deviations are shown for c_{11} and c_{33}. The error estimation for c_{13} was obtained from Tables IV and V.

Effects of formalin fixation on the ultrasonic properties of soft tissue have been of interest for some time (Bamber *et al.*, 1979; Hall *et al.*, 2000; Shung and Reid, 1978; van der Steen *et al.*, 1991). The current study permits a direct comparison of the stiffness coefficients of freshly excised and formalin-fixed myocardium. The increases in the values of elastic coefficients, which are clearly illustrated in Fig. 3, are consistent with the observation that the tissue feels noticeably stiffer after formalin fixation. The ratios of c_{33}/c_{11} in myocardium also changed after formalin fixation, which indicates a different degree of anisotropy for the elastic properties. A careful examination of the results summarized in Fig. 3 indicates that the effects of formalin fixation on elastic stiffness coefficients are larger along the myofibers than perpendicular to the myofibers. It is interesting that the same

trend has been observed in tendon (Hoffmeister *et al.*, 1996b; Kuo *et al.*, 2001). Compared with myocardium [~1%–2% collagen by wet weight (Goss *et al.*, 1980)], tendon has a larger concentration of collagen [~30% collagen by wet weight (Goss *et al.*, 1980)] and a higher degree of alignment. Kuo *et al.* measured the elastic coefficients c_{11}, c_{33}, and c_{13} for fresh tendon and compared results with those for formalin-fixed tendon reported by our laboratory. The elastic coefficients for fresh tendon were smaller than, but of the same order of magnitude as, those for formalin-fixed tendon (Kuo *et al.*, 2001). Kuo *et al.* also showed that there was a larger difference for fresh than for formalin-fixed tendon in values of velocity along the fiber axis and therefore in c_{33}.

One may obtain an assessment of the degree of anisotropy in myocardium from Tables II and III. If either transversely isotropic or orthortropic symmetry is assumed, the ratios of $(c_{33}-c_{11})/c_{33}$ are no more than 2% for both freshly excised and formalin-fixed myocardium. Although this might initially seem to justify the use of an isotropic model for the mechanical properties of the heart, studies from our laboratory indicate that a modest anisotropy in the stiffness can play a significant role in the angular dependence of Young's modulus (Trousil *et al.*, 2003). To better understand the mechanical properties of myocardium, symmetries described by more than two distinct coefficients (i.e., isotropic symmetry) might therefore be necessary. If transversely isotropic symmetry is assumed, where five distinct elastic stiffness coefficients are required to describe elastic properties, measurements of the transverse ultrasonic velocities parallel and perpendicular to the myofibers would also be needed. For orthotropic symmetry, the wave propagation characteristics are summarized in Table VI for the evaluation of elastic coefficients from phase velocity measurements (Auld, 1990). As indicated in Table VI, measurements in additional planes would be required to obtain all nine distinct stiffness coefficients. As shown earlier, the stiffness coefficients we report here (c_{11}, c_{33}, and c_{13}) are the same for both transversely isotropic and orthotropic symmetries. Although other studies (Chen *et al.*, 2005) considered the possible role of the sheet structure of the heart, more work is required to explore the consequences of such an approach on ultrasonic measurements.

Because myocardium is a weakly dispersive medium, the stiffness coefficients also exhibit modest frequency dependences. However, for the bandwidth of the current measurements (3 to 8 MHz), the variation did not have a significant impact on the stiffness coefficients. For comparison, the same techniques as described in the previous sections were applied for broadband (3 to 8 MHz) measurements of time-of-flight velocity as a function of angle of insonification (Baldwin *et al.*, 2005b). Results showed $c_{11}=2.565$ GPa, $c_{33}=2.592$ GPa, and $c_{13}=2.568$ GPa for freshly excised myocardium and $c_{11}=2.613$ GPa, $c_{33}=2.653$ GPa, and $c_{13}=2.616$ GPa for formalin-fixed myocardium. Comparison of these values with those listed in Tables II–V indicates that the weak dispersion does not play a significant role in the determination of these elastic stiffness coefficients in the bandwidth of the current study. The frequency bandwidth in this study is in the range of that typically used in the clinical setting, namely adult echocardiography and pediatric echocardiography. From the existing literature, Kanai (2005) summarized transverse mode elastic values for soft tissue, including myocardium, over a much wider frequency range. The frequency dependence of stiffness coefficients is of considerable interest. Additional work is required to investigate this issue. For example, a viscoelastic model would presumably result in an explicit frequency dependence.

TABLE IV. Influence of c_{44} on c_{13} for freshly excised myocardium. Results from representative frequencies 3, 5, and 7 MHz are shown.

		Likely values					Less likely values
c_{44} (MPa)		0	0.01	0.1	1	10	100
V_{shear} (m/s)		0	~3	~10	~30	~100	~300
c_{13} (MPa)	3 MHz	2.562	2.562	2.562	2.560	2.542	2.362
	5 MHz	2.564	2.564	2.564	2.562	2.544	2.364
	7 MHz	2.566	2.566	2.566	2.564	2.546	2.366

TABLE V. Influence of c_{44} on c_{13} for formalin-fixed myocardium. Results from representative frequencies 3, 5, and 7 MHz are shown.

		Likely values					Less likely values
c_{44} (MPa)		0	0.01	0.1	1	10	100
V_{shear} (m/s)		0	~3	~10	~30	~100	~300
c_{13} (MPa)	3 MHz	2.608	2.608	2.608	2.606	2.590	2.408
	5 MHz	2.612	2.612	2.612	2.610	2.594	2.412
	7 MHz	2.615	2.615	2.615	2.613	2.595	2.415

V. SUMMARY

This study reported the elastic stiffness coefficients c_{11}, c_{33}, and c_{13} for freshly excised and formalin-fixed myocardium for both orthotropic and transversely isotropic symmetry models. Only a very modest frequency dependence of the elastic stiffness coefficients was present for the frequency range in the current measurements (3 to 8 MHz). Comparison of those coefficients for freshly excised and formalin-fixed myocardium showed increases for the stiffness coefficients after formalin fixation and also showed that the effect of fixation was greater parallel to the myofibers than perpendicular to the myofibers.

ACKNOWLEDGMENT

This work was supported in part by NIH Grant No. R37 HL40302.

TABLE VI. Elastic stiffness coefficients that determine the phase velocity of the waves propagating in specific planes for orthotropic symmetry.

	Velocity of shear wave ↓	Velocity of quasi-longitudinal wave ↓	Velocity of quasi-shear wave ↓
Plane 13	c_{44}, c_{66}	c_{11}, c_{33}, c_{13}, c_{55}	c_{11}, c_{33}, c_{13}, c_{55}
Plane 23	c_{55}, c_{66}	c_{22}, c_{33}, c_{23}, c_{44}	c_{22}, c_{33}, c_{23}, c_{44}
Plane 12	c_{44}, c_{55}	c_{11}, c_{22}, c_{12}, c_{66}	c_{11}, c_{22}, c_{12}, c_{66}

APPENDIX: BRIEF REVIEW OF LINEAR ELASTIC THEORY

For a linear elastic lossless medium, Hooke's law states that the stress is a linear function of all the strain components (Auld, 1990),

$$T_{ij} = \sum_{k=1}^{3} \sum_{l=1}^{3} c_{ijkl} S_{kl} \tag{A1}$$

where $i,j,k,l=1,2,3$ (representing coordinate axes 1,2,3) and where the 81 elastic stiffness coefficients c_{ijkl} fully characterize the elastic properties of the material. Because of the symmetry of stress and strain, the general elastic constitutive relations can be represented by the reduced-subscript (Voigt) notation,

$$T_I = \sum_{J=1}^{6} c_{IJ} S_J, \tag{A2}$$

where $I,J=1,2,3,4,5,6$ and c_{IJ} becomes a 6×6 stiffness matrix. In addition, the requirement of positive definite strain energy reduces the maximum number of independent stiffness coefficients to 21. Certain constraints (physical realizability conditions) on the c_{IJ} matrix must be satisfied as a direct result of positive definite strain energy. The number of independent terms in the c_{IJ} matrix may be further reduced by the symmetry of the medium. Orthotropic symmetry leads to nine distinct stiffness coefficients (the same as for an orthorhombic crystal), whereas transversely isotropic symmetry leads to five distinct stiffness coefficients (the same as for a hexagonal crystal).

For orthotropic symmetry, there exist three orthogonal planes of mirror symmetry [Plane_12, Plane_13 and Plane_23 as shown in Fig. 1(a)]. The nine distinct coefficients required to describe the elastic properties of a medium with this symmetry are shown in the following matrix:

$$[c_{IJ}^{\mathrm{Ortho}}] = \begin{bmatrix} c_{11}^{\mathrm{Ortho}} & c_{12}^{\mathrm{Ortho}} & c_{13}^{\mathrm{Ortho}} & 0 & 0 & 0 \\ c_{12}^{\mathrm{Ortho}} & c_{22}^{\mathrm{Ortho}} & c_{23}^{\mathrm{Ortho}} & 0 & 0 & 0 \\ c_{13}^{\mathrm{Ortho}} & c_{23}^{\mathrm{Ortho}} & c_{33}^{\mathrm{Ortho}} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44}^{\mathrm{Ortho}} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{55}^{\mathrm{Ortho}} & 0 \\ 0 & 0 & 0 & 0 & 0 & c_{66}^{\mathrm{Ortho}} \end{bmatrix}. \tag{A3}$$

A transversely isotropic medium exhibits a higher symmetry and therefore requires fewer distinct elastic coefficients. This is illustrated in Fig. 1(b) which depicts a Cartesian coordinate system in which the 3-axis represents the one distinct direction and the other two axes (1 and 2) are equivalent to each other. The elastic properties of the medium can then be described with five distinct coefficients as expressed by the following matrix:

$$[c_{IJ}^{\mathrm{TIso}}] = \begin{bmatrix} c_{11}^{\mathrm{TIso}} & c_{11}^{\mathrm{TIso}} - 2c_{66}^{\mathrm{TIso}} & c_{13}^{\mathrm{TIso}} & 0 & 0 & 0 \\ c_{11}^{\mathrm{TIso}} - 2c_{66}^{\mathrm{TIso}} & c_{11}^{\mathrm{TIso}} & c_{13}^{\mathrm{TIso}} & 0 & 0 & 0 \\ c_{13}^{\mathrm{TIso}} & c_{13}^{\mathrm{TIso}} & c_{33}^{\mathrm{TIso}} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44}^{\mathrm{TIso}} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44}^{\mathrm{TIso}} & 0 \\ 0 & 0 & 0 & 0 & 0 & c_{66}^{\mathrm{TIso}} \end{bmatrix}. \tag{A4}$$

In the case of highest symmetry, that of isotropy, the elastic matrix contains only two independent coefficients. The corresponding stiffness matrix can be expressed with the elastic coefficients c_{11}^{Iso} and c_{44}^{Iso}

$$[c_{IJ}^{\mathrm{Iso}}] = \begin{bmatrix} c_{11}^{\mathrm{Iso}} & c_{11}^{\mathrm{Iso}} - 2c_{44}^{\mathrm{Iso}} & c_{11}^{\mathrm{Iso}} - 2c_{44}^{\mathrm{Iso}} & 0 & 0 & 0 \\ c_{11}^{\mathrm{Iso}} - 2c_{44}^{\mathrm{Iso}} & c_{11}^{\mathrm{TIso}} & c_{11}^{\mathrm{Iso}} - 2c_{44}^{\mathrm{Iso}} & 0 & 0 & 0 \\ c_{11}^{\mathrm{Iso}} - 2c_{44}^{\mathrm{Iso}} & c_{11}^{\mathrm{Iso}} - 2c_{44}^{\mathrm{Iso}} & c_{11}^{\mathrm{Iso}} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44}^{\mathrm{Iso}} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44}^{\mathrm{Iso}} & 0 \\ 0 & 0 & 0 & 0 & 0 & c_{44}^{\mathrm{Iso}} \end{bmatrix} \tag{A5}$$

or the widely used Lamé constants λ and μ,

$$[c_{IJ}^{\mathrm{Iso}}] = \begin{bmatrix} \lambda+2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda+2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda+2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{bmatrix}. \tag{A6}$$

The relationships among the stiffness matrices in Eqs. (A3)–(A5) are summarized in Table I.

Auld, B. A. (**1990**). *Acoustic Fields and Waves in Solids* (Krieger, Malabar, FL).

Baldwin, S. L., Marutyan, K. R., Yang, M., Wallace, K. D., Holland, M. R., and Miller, J. G. (**2005a**). "Estimating myocardial attenuation properties from m-mode ultrasonic backscatter," Ultrasound Med. Biol. **31**, 477–484.

Baldwin, S. L., Yang, M., Marutyan, K. R., Wallace, K. D., Holland, M. R., and Miller, J. G. (**2005b**). "Measurement of the anisotropy of ultrasonic velocity in freshly excised and formalin-fixed myocardial tissue," J. Acoust. Soc. Am. **118**, 505–513.

Bamber, J. C., Hill, C. R., King, J. A., and Dunn, F. (**1979**). "Ultrasonic propagation through fixed and unfixed tissues," Ultrasound Med. Biol. **5**, 159–165.

Chen, J., Liu, W., Zhang, H., Lacy, L., Yang, X., Song, S. K., Wickline, S. A., and Yu, X. (**2005**). "Regional ventricular wall thickening reflects changes in cardiac fiber and sheet structure during contraction: Quantification with diffusion tensor mri," Am. J. Physiol. Heart Circ. Physiol. **289**, H1898–H1907.

D'Hooge, J., Heimdal, A., Jamal, F., Kukulski, T., Bijnens, B., Rademakers, F., Hatle, L., Suetens, P., and Sutherland, G. R. (**2000**). "Regional strain and strain rate measurements by cardiac ultrasound: Principles, implementation and limitations," Eur. J. Echocardiogr. **1**, 154–170.

D'Hooge, J., Konofagou, E., Jamal, F., Heimdal, A., Barrios, L., Bijnens, B., Thoen, J., Van de Werf, F., Sutherland, G., and Suetens, P. (**2002**). "Two-dimensional ultrasonic strain rate measurement of the human heart in

vivo," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **49**, 281–286.
Del Grosso, V. A., and Mader, C. W. (**1972**). "Speed of sound in pure water," J. Acoust. Soc. Am. **52**, 1442–1446.
Eroglu, E., Herbots, L., Van Cleemput, J., Droogne, W., Claus, P., D'Hooge, J., Bijnens, B., Vanhaecke, J., and Sutherland, G. R. (**2005**). "Ultrasonic strain/strain rate imaging—a new clinical tool to evaluate the transplanted heart," Eur. J. Echocardiogr. **6**, 186–195.
Frizzell, L. A., and Carstensen, E. L. (**1976**). "Shear properties of mammalian tissues at low megahertz frequencies," J. Acoust. Soc. Am. **60**, 1409–1411.
Gilman, G., Khandheria, B. K., Hagen, M. E., Abraham, T. P., Seward, J. B., and Belohlavek, M. (**2004**). "Strain rate and strain: A step-by-step approach to image and data acquisition," J. Am. Soc. Echocardiogr **17**, 1011–1020.
Goss, S. A., Frizzell, L. A., Dunn, F., and Dines, K. A., (**1980**). "Dependence of the ultrasonic properties of biological tissue on constituent proteins," J. Acoust. Soc. Am. **67**, 1041–1044.
Hall, C. S., Dent, C. L., Scott, M. J., and Wickline, S. A. (**2000**). "High-frequency ultrasound detection of the temporal evolution of protein cross linking in myocardial tissue," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **47**, 1051–1058.
Hoffmeister, B. K., Gehr, S. E., and Miller, J. G. (**1996a**). "Anisotropy of the transverse mode ultrasonic properties of fixed tendon and fixed myocardium," J. Acoust. Soc. Am. **99**, 3826–3836.
Hoffmeister, B. K., Handley, S. M., Wickline, S. A., and Miller, J. G. (**1996b**). "Ultrasonic determination of the anisotropy of Young's modulus of fixed tendon and fixed myocardium," J. Acoust. Soc. Am. **100**, 3933–3940.
Hoffmeister, B. K., Handley, S. M., Verdonk, E. D., Wickline, S. A., and Miller, J. G. (**1995**). "Estimation of the elastic stiffness coefficient c_{13} of fixed tendon and fixed myocardium," J. Acoust. Soc. Am. **97**, 3171–3176.
Humphery, J. D. (**2001**). *Cardiovascular Solid Mechanics: Cells, Tissues, and Organs* (Springer-Verlag, New York).
Inaba, Y., Yuda, S., Kobayashi, N., Hashimoto, A., Uno, K., Nakata, T., Tsuchihashi, K., Miura, T., Ura, N., and Shimamoto, K. (**2005**). "Strain rate imaging for noninvasive functional quantification of the left atrium: Comparative studies in controls and patients with a trial fibrillation," J. Am. Soc. Echocardiogr **18**, 729–736.
Kanai, H. (**2005**). "Propagation of spontaneously actuated pulsive vibration in human heart wall and in vivo viscoelasticity estimation," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **52**, 1931–1942.
Kuo, P. L., Li, P. C., and Li, M. L. (**2001**). "Elastic properties of tendon measured by two different approaches," Ultrasound Med. Biol. **27**, 1275–1284.
LeGrice, I. J., Smaill, B. H., Chai, L. Z., Edgar, S. G., Gavin, J. B., and Hunter, P. J. (**1995**). "Laminar structure of the heart-ventricular myocyte arrangement and connective-tissue architecture in the dog," Am. J. Physiol. Heart Circ. Physiol. **38**, H571–H582.
Madsen, E. L., Sathoff, H. J., and Zagzebski, J. A. (**1983**). "Ultrasonic shear wave properties of soft tissues and tissuelike materials," J. Acoust. Soc. Am. **74**, 1346–1355.
Mottley, J. G., and Miller, J. G. (**1990**). "Anisotropy of the ultrasonic attenuation in soft tissues: Measurements *in vitro*," J. Acoust. Soc. Am. **88**, 1203–1210.
O'Donnell, M., and Skovoroda, A. R. (**2004**). "Prospects for elasticity reconstruction in the heart," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **51**, 322–328.
Shung, K. K., and Reid, J. M. (**1978**). "Ultrasound velocity in major bovine blood vessel walls," J. Acoust. Soc. Am. **64**, 692–694.
Streeter, D. D., Jr., and Hanna, W. T. (**1973**). "Engineering mechanics for successive states in canine left ventricular myocardium. II. Fiber angle and sarcomere length," Circ. Res. **33**, 656–664.
Sutherland, G. R., Di Salvo, G., Claus, P., D'Hooge, J., and Bijnens, B. (**2004**). "Strain and strain rate imaging: A new clinical approach to quantifying regional myocardial function," J. Am. Soc. Echocardiogr **17**, 788–802.
Trousil, R. L., Waters, K. R., and Miller, J. G. (**2001**). "Experimental validation of the use of kramers-kronig relations to eliminate the phase sheet ambiguity in broadband phase spectroscopy," J. Acoust. Soc. Am. **109**, 2236–2244.
Trousil, R. L., Handley, S. M., Wallace, K. D., Holland, M. R., and Miller, J. G. (**2003**). "Improving strain-derived estimates of regional cardiac performance by accounting for myocardial anisotropy," IEEE Ultrason. Symp. Proc. 03CH37476C, pp. 138–141.
Usyk, T. P., Mazhari, R., and McCulloch, A. D. (**2000**). "Effect of laminar orthotropic myofiber architecture on regional stress and strain in the canine left ventricle," J. Elast. **61**, 143–164.
van der Steen, A. F., Cuypers, M. H., Thijssen, J. M., and de Wilde, P. C. (**1991**). "Influence of histochemical preparation on acoustic parameters of liver tissue: A 5-mhz study," Ultrasound Med. Biol. **17**, 879–891.
Verdonk, E. D., Wickline, S. A., and Miller, J. G. (**1992**). "Anisotropy of ultrasonic velocity and elastic properties in normal human myocardium," J. Acoust. Soc. Am. **92**, 3039–3050.
Verdonk, E. D., Hoffmeister, B. K., Wickline, S. A., and Miller, J. G. (**1996**). "Anisotropy of the slope of ultrasonic attenuation in formalin fixed human myocardium," J. Acoust. Soc. Am. **99**, 3837–3843.
Weidemann, F., Dommke, C., Bijnens, B., Claus, P., D'Hooge, J., Mertens, P., Verbeken, E., Maes, A., Van de Werf, F., De Scheerder, I., and Sutherland, G. R. (**2003**). "Defining the transmurality of a chronic myocardial infarction by ultrasonic strain-rate imaging: Implications for identifying intramural viability: An experimental study," Circulation **107**, 883–888.
Williams, R. I., Payne, N., Phillips, T., D'Hooge, J., and Fraser, A. G. (**2005**). "Strain rate imaging after dynamic stress provides objective evidence of persistent regional myocardial dysfunction in ischaemic myocardium: Regional stunning identified?" Heart **91**, 152–160.
Yuda, S., Short, L., Leano, R., and Marwick, T. H. (**2002**). "Myocardial abnormalities in hypertensive patients with normal and abnormal left ventricular filling: A study of ultrasound tissue characterization and strain," Clin. Sci. **103**, 283–293.

Erratum: "Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners" [J. Acoust. Soc. Am. 116(4), 2395–2405 (2004)]

Mead C. Killion, Patricia A. Niquette, and Gail I. Gudmundsen
Etymotic Research, Inc., 61 Martin Lane, Elk Grove Village, Illinois 60007

Lawrence J. Revit[a)]
Revitronix, Brownsville, Vermont 05037

Shilpi Banerjee[b)]
Northwestern University, Evanston, Illinois 60208

(Received 20 December 2005; accepted 22 December 2005)

[DOI: 10.1121/1.2166610]

PACS number(s): 43.71.Ky, 43.71.Gv, 43.72.Dv, 43.10.Vx [KWG]

In the original Tables I and II, the numbers for the 80% confidence intervals and 80% critical differences were incorrectly calculated as 1.6* Standard Deviation (corresponding to 90%) rather than the correct 1.28* Standard Deviation (80%). The revised Tables I and II show the correct values for all three confidence levels (80%, 90%, and 95%).

TABLE I. Reliability: number of QuickSIN lists required for a given accuracy. An 80% confidence interval is normally adequate for clinical testing, where the results of any one test are used in context with other factors. A single QuickSIN list is accurate to ±1.8 dB at the 80% confidence interval, and to ±2.2 dB at the 90% confidence interval. A 95% confidence interval is common for research reporting, where a reduced risk of error is normally required. In this case, a sngle QuickSIN list is accurate to ±2.7 dB at the 95% confidence interval.

	Number of lists								
	1	2	3	4	5	6	7	8	9
95% C.I. ±, in dB	2.7	1.9	1.6	1.4	1.2	1.1	1.0	1.0	0.9
90% C.I. ±, in dB	2.2	1.6	1.3	1.1	1.0	0.9	0.8	0.8	0.7
80% C.I. ±, in dB	1.8	1.3	1.0	0.9	0.8	0.7	0.7	0.6	0.6

TABLE II. QuickSIN critical differences for comparing two conditions (e.g., two hearing aids or hearing aid adjustments). If one list per condition is used, results must be greater than 2.5 dB different to be considered statistically significant at the 80% confidence interval, 3.2 dB at the 90% confidence interval, and 3.9 dB at the 95% confidence interval.

	Lists per condition								
	1	2	3	4	5	6	7	8	9
95% C.D. ±, in dB	3.9	2.7	2.2	1.9	1.7	1.6	1.5	1.4	1.3
90% C.D. ±, in dB	3.2	2.2	1.8	1.6	1.4	1.3	1.2	1.1	1.1
80% C.D. ±, in dB	2.5	1.8	1.5	1.3	1.1	1.0	1.0	0.9	0.8

[a)]Current address: Revitronix, Braintree, VT 05060.
[b)]Current address: Starkey Laboratories, Eden Prairie, MN 55344.

 0001-4966/2006/119(3)/1888/1/$22.50